ELSEVIER

Theoretical Computer Science 172 (1997) 91–120

Theoretical Computer Science

PAC learning of concept classes through the boundaries of their items

B. Apolloni *, S. Chiaravalli

Department of Computer Science, University of Milan, Via Comelico 39-41, 20135 Milan, Italy

Received August 1994
Communicated by J. Becvár

Abstract

We present a new perspective for investigating the probably approximate correct (PAC) learnability of classes of concepts. We focus on special sets of points for characterizing the concepts within their class. This gives rise to a general notion of boundary of a concept, which holds even in discrete spaces, and to a special probability measuring technique. This technique is applied (i) to narrow the gap between the minimum and maximum sample sizes necessary to learn even under a more stringent learnability definition, and (ii) to get self-explanatory indices of the complexity of the learning task. These indices can be roughly estimated during the learning process and appear very useful in the treatment of nonsymbolic procedures, e.g. in the context of neutral networks.

1. Introduction

Learning – in the sense of building up of a routine not yet available in our software library, on the sole basis of a set of examples about how the routine has to behave – is often a costly job for computing machineries. The probably approximately correct (PAC) model [11], introduced by Valiant in 1984, supplies a natural framework for evaluating the complexity of machine learning procedures. The main ingredients of PAC learning are:

- a probability space $(X, \mathscr{F}, P)$, where
  X is the set of the possible outcomes of a source of random data,
  $\mathscr{F}$ is a σ-algebra on X, and
  P is a (possibly unknown) probability measure defined over $\mathscr{F}$;
- a set $\mathbf{C}$ of subsets c of X belonging to $\mathscr{F}$. Every c is called a *concept*, and $\mathbf{C}$ is a *concept class*;

* Corresponding author.

SSDI 0304-3975(95)00240-5

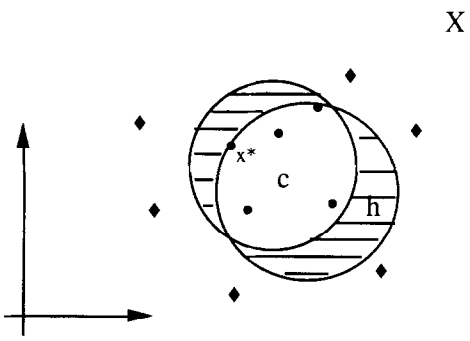

Fig. 1. A PAC learning framework. (X is the set of points belonging to the cartesian plane, c a concept from the concept class of circles; h a hypothesis from the same concept class, (•) = 1-labelled (positive) sampled points; (♦) = 0-labelled (negative) sampled points).

- a labelled random sample ξ_m^c, which, for each c, is constituted of pairs $\{(\xi_i, \chi_c(\xi_i)), i = 1,\ldots,m\}$, where $\xi_1,\ldots,\xi_m$ are randomly chosen from X and $\chi_c : X \mapsto \{0,1\}$ is the *characteristic function* of c; thus, by definition, $\chi_c(\xi) = 1$ if and only if ξ is an element of c.

In strictly mathematical terms, learning consists in computing a statistics on ξ_m^c. The output of this computation is a symbolical description of a region $h \subseteq X, h \in \mathcal{F}$, which we call *hypothesis* and assume as an *estimate* of c. The probability value $\varepsilon = P(c \div h)$ of the symmetric difference between c and h is assumed to be the *loss function* [15] of our estimate, thus providing an accuracy parameter for the hypothesis. A learning algorithm is a procedure **A** to generate a family of hypotheses h_m with their respective parameters ε_m converging to 0 in probability.

For instance, in Fig. 1 the set X of the experimental outcomes coincides with the cartesian plane; the learning task is to identify one particular circle c within the concept class **C** of all possible circles in the cartesian plane. This might be a mathematical model for identifying the site and the emission range of a source of radiating pollution, such as noise, X-ray and so on, in a flat homogeneous region. In this case it is reasonable to regard circles as possible hypotheses, so that also the set of hypotheses is represented by **C**. Our labelled sample might be identified with a set of randomly distributed monitoring stations. The ith station is completely described by its position ξ_i in the plane, together with $\{0,1\}$-valued variable, telling us whether pollution is detected – above a given threshold – by the station. We are concerned with the probability that Mr. John Smith is exposed to radiation, assuming the population to have the same distribution as the set of monitoring stations. Thus, the accuracy of the hypothesis refers not directly to the portion of region which is misclassified – the part which is subjected to pollution but the authority declares safe on the basis of the above monitorings and viceversa – rather to the probability that Mr. Smith lives in this region. With a few minor caveats, to be discussed later, our theory is applicable to any concept class. In the course of our exposition, to fix the ideas, we shall indulge on discussing some simple and clearly defined examples.

Disregarding the most ambitious goal of perfect ($\varepsilon = 0$) learning, which is only an asymptotical target of other theories [8], in Valiant approach [11] one is able to compute stringent bounds on the sample size needed to achieve any preassigned accuracy and confidence to the learning algorithm. These bounds are connected to some combinatorial indices of complexity of **C**.

With respect to other well-known statistical methods for domain estimate, such as those given by equivalent statistical blocks [9, 10] or by confidence regions [15], a distinguishing feature of PAC learning estimate is the restriction to regions h of a certain prescribed shape: this has strong impact on the informative content of the sampled points.

Vapnik [12] and Valiant [5] use the sampled points as witnesses of the concept c to be learned. Accordingly, if a 1-labelled point is observed in a given position of the sample space, it means that this position is included in the concept and vice-versa for 0-labelled points. Moreover, owing to our assumptions about the class to which c belongs, some of these points play the role of markers which extend their label to specially shaped surrounding regions. For example, for the concept and hypothesis classes of Fig. 1, surroundings have the shape of waning or waxing moon, like those marked in this picture.

From a different perspective, using a military metaphor just as a joke, we look at the sample points as sentinels along a frontier. Namely, consider the enemy region constituted by the symmetric difference $c \div h$. We call sentinels those (positive, as well as negative) sampled points that forbid the expansion of $c \div h$. For instance in Fig. 1 point x^* is a sentinel if c is fixed and $c \div h$ might grow only by inclusion. The remaining sampled points constitute then rear-guard which, if numerous and fairly scattered on X, gives confidence that each possible critical point of the frontier has been considered in our list of the guard points.

Fix **C** and consider the task of watching on the concepts of **C**. Depending on our watching strategy, for any concept c there will exist a minimum number $n(c)$ of sentinels for watching. Let d_C be the supremum of all $n(c)$'s as c ranges over all concepts of **C**. We assume d_C as an index of complexity for **C** and we call it detail of **C**.

From a more statistical point of view, the confidence δ on the measure ε of a watched region arises from the comparison between the number of sampled points and the number of sentinels which are necessary to watch the region.

In the line of rank order statistics theory [9, 10], the key idea is the following:

Let $R_1 \subseteq R_2 \subseteq R_3 \subseteq \cdots$ be the set of growing subsets $c \div h$'s depicted in Fig. 2. Whatever the probability distribution P on X, for any two such regions R' and R'' we have that $R' \subseteq R'' \Leftrightarrow P(R') \leqslant P(R'')$. Moreover, assume that an algorithm A computes statistics from random samples of size m drawn by X. In particular, the output of A on sampled points $\xi_1, \xi_2, \ldots, \xi_m$, is a symbolic description of the region R.

Now, let us identify event E with its description, namely, $E = [\text{description of } E]$. By R_α we mean a subset R_i such that $P(R_i) = \alpha$, let $E_1 = [R_\alpha \text{ contains the sentinels of } R]$, $E_2 = [R \subseteq R_\alpha]$ and $E_3 = [P(R) \leqslant \alpha]$. Then, after minor cautions, E_1 implies both E_2 and E_3. Consequently, letting $X^{(m)}$ be the cartesian product of m copies of X and $P^{(m)}$ be the

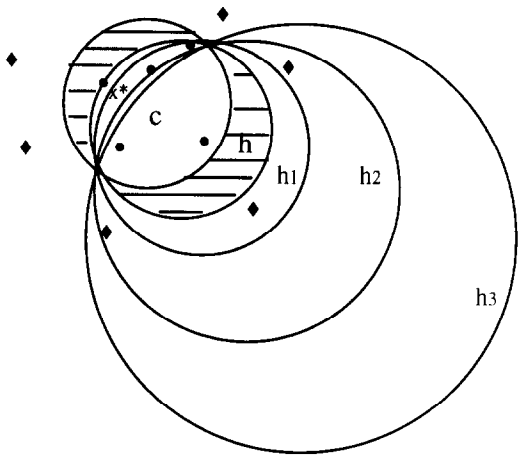

Fig. 2. Regions growing by inclusion. $R = c \div h; R_i = c \div h_i$. If R_j contains the sentinels of R, then $R \subseteq R_j$, and $P(R) \leqslant P(R_j)$

measure of the m-fold product probability space $(X^{(m)}, \mathscr{F}^{(m)}, P^{(m)}), P^{(m)}([P(R) \leqslant \alpha]) \geqslant P^{(m)}(E1)$.

If on each sample R has exactly k sentinels, for a given k, a special selection of the above regions R_i will allow us to obtain that E_1 is equivalent to the event $E_4 =$ [at least k sampled points fall in R_α], so that $P^{(m)}(E_1)$ is computed from the binomial distribution law. Finally, if the sentinels are in number less or equal k, then the above probability is a suitable lower bound on $P^{(m)}(P(R) \leqslant \alpha)$.

For the validity of the above argument, one must guarantee that the error measure is a random variable whose infimum equals 0. This requires stating prescriptions on the "completeness" of the learning algorithm in assigning hypotheses to samples. Therefore,

(i) we characterize this assignment in terms of the number $n(c)$ of arguments which are necessary to it for pointing exactly to the concept c which has to be learnt. We call *arity* of the learning algorithm the supremum of $n(c)$, where c ranges over $\mathbf{C}$.

(ii) we establish a link between the arity of the learning algorithm and the learnability of the concept class.

Some other complexity indices are proposed in the literature, such as Vapnik–Chervonenkis dimensions [13] or width [6]. Among the main distinguishing features of our approach let us mention the following:

1. Our notion of detail and arity is more directly linked to the complexity of actual learning procedures.

2. By seeing concepts through these indices and using the above probability measuring technique, we are allowed to tighten the usual gap between lower and upper bounds on sample sizes needed for obtaining prescribed learning accuracies.

3. Detail and arity can be roughly estimated already during the learning process, just in connection with that portion of the hypothesis class which is effectively spanned.

Moreover, without any attempt to lower the complexity of the whole learning task, we highlight a possible nonuniformity of the complexity of the hypothesis class with

respect to the accuracy target which occurs in subsymbolic learning [7]. This might motivate the general lower cost of the related learning paradigms.

2. Bordering the concepts

Suppose we are given:
– a nonempty set X
– a $\{0,1\}$-valued function $b : X \mapsto \{0,1\}$, which we shall call *Boolean* function over X
– a vector of pairs $((x_i, b(x_i)), i = 1,\ldots,n)$, with $x_i \in X$.
By abuse of notation, we shall not make any distinction between b and its support, i.e. the set of elements x such that $b(x_i) = 1$. We call the pairs $(x_i, b(x_i))$ *labelled points* for giving a geometrical evidence to the matter and the set of pairs *sample* for meaning that these points might be whichever in X, some possibly coinciding, though they do not constitute, here, a random sample.

Now, imagine stating an inference rule **A** for discovering b from the above labelled points. A minimal requisite to be satisfied by any rule is *consistency*: for each sample, the inferred function must compute well at least on the sample points.

Then, we are interested in selecting from the sample the pivots of this inference, i.e. that minimal subset of "sentinels" which alone imply the consistency of the hypothesis. A first result of this section is that under very general conditions these pivots are shattered [3] by the image of **A**.

The remaining of this section is aimed at stating further connections between the sentinels and the whole sample which makes **A** inferring b.

We are only concerned with algorithmic inference rules – we don't allow such devices as oracles, random bit generators and the like. Accordingly, we start by defining a *hypothesis class* H as a family of subsets of X arising from the output of some function **F** of the samples taken from X. **F** will represent the core of any **A**, as will be shown later. The first result on the pivots of the inference rule is obtained by studying H as a general collection **C** of subsets of X. The remaining connections between pivots and whole sample will concern a restriction Δ of **F**.

Notational conventions:
– Sets will be denoted by both small and capital letters, with the tendency to use lower-case letter (e.g. x) for meaning an element of a set denoted by capital of the same letter (e.g. X).
– Bold capital letter shall be usually deserved to denote set of sets (e.g. **C**). Bold lower case letter shall denote vectors (e.g. $\boldsymbol{x}$).
– For any vector $\boldsymbol{x}$, $|\boldsymbol{x}|$ denotes the length of $\boldsymbol{x}$, i.e. the number of its components, set($\boldsymbol{x}$)denotes the set $\{x_i\}$ of all the $\boldsymbol{x}$ occurring among the components of $\boldsymbol{x}$.
– By $\#B$ we shall denote the cardinality of B; $\overline{B}$ shall denote the complement of B, when the universe is clear from the context. As usual, $\cap$ and $\cup$ shall denote set-theoretic

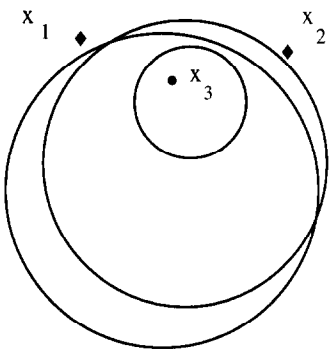

Fig. 3. **F** associates three subsets of the cartesian plane to the vector (x_1, x_2, x_3) when its components are partitioned in two external points: x_1 and x_2 and one inner point: x_3.

intersections and unions, respectively. Following tradition, the symbol $\neg$ shall denote negation.

– $(y \mid \phi(y))$ means "y such that $\varphi(y)$ is true". Note that we will never use $\{ \mid \}$. We will use $\{(\mid)\}$ sometimes with obvious meaning.

Definition 1. Given a set X and a vector $\boldsymbol{x} = (x_1, \ldots, x_m)$ over X (in the sense that all components x_i of $\boldsymbol{x}$ are elements of X), by a partitioned *vector* we mean a pair $(\boldsymbol{x}, \pi)$ where $\pi \subseteq \{1, \ldots, |\boldsymbol{x}|\}$ and the following condition is satisfied:

(i) (Noncontradiction) Whenever $i, j \in \{1, \ldots, |\boldsymbol{x}|\}$ and $x_i = x_j$, then $i \in \pi$ iff $j \in \pi$.

Let W be the *set of partitioned vectors* over X. Then *contouring function* is a map $\mathbf{F} : (\boldsymbol{x}, \pi) \mapsto \mathbf{F}(\boldsymbol{x}, \pi) \subset 2^{\mathbf{x}}$ whose domain is a subset W' of W obeying to the following constraint:

(ii) (Completeness) For every vector $\boldsymbol{x}$ of any length over X, there exists at least one partition π such that $\mathbf{F}(\boldsymbol{x}, \pi)$ is defined.

The map assigns to every partitioned vector $(\boldsymbol{x}, \pi) \in W'$ as set $\mathbf{F}(\boldsymbol{x}, \pi)$ of subsets of X with the following property:

(iii) (Consistency) Whenever $(\boldsymbol{x}, \pi)$ is in the domain of $\mathbf{F}$ and $c \in \mathbf{F}(\boldsymbol{x}, \pi)$ then $i \in \pi$ iff $x_i \in c$. We mean by $\mathbf{H}$ the union of all sets $\mathbf{F}(\boldsymbol{x}, \pi)$ where $(\boldsymbol{x}, \pi)$ ranges over the domain of definition of $\mathbf{F}$. Thus, a set $c \subseteq X$ belongs to $\mathbf{H}$ iff for some partitioned vector $(\boldsymbol{x}, \pi)$, $c \in \mathbf{F}(\boldsymbol{x}, \pi)$.

If in this definition we omit the consistency condition (iii), we obtain a notion of *precontouring function*. In the rest of paper, by $\mathbf{G}$ we shall always denote a precontouring function.

Example 1. Given $X = \mathbb{R}^2$, an instance $\mathbf{F}((x_1, x_2, x_3), \{3\})$ of a possible $\mathbf{F}$ for the set $\mathbf{H}$ of circles on the cartesian plane is shown in Fig. 3.

Remark 1. We insist that $\mathbf{F}$ associates sets of subsets of X to vectors of labelled points. In the following we will assume that these points are randomly sampled from

X, and will focus on the size of this sample. In particular:

(a) the same item might fill more locations of the vector;

(b) the partition π assigns labels to points;

(c) these labels are to signify membership to a some subset a of X, which belongs to $\mathbf{H}$ if condition (iii) holds; namely points whose indices are in π are labelled by "+" or "1" ($\bullet$ in the pictures) belong to a, the other points are labelled by "−" or "0" ($\blacklozenge$ in the pictures) with the complementary meaning.

(d) $\mathbf{F}(\boldsymbol{x}, \pi)$ lists a set of subsets h of $\mathbf{H}$, to which the above labelling is true. Condition (ii) is only a prerequisite for having a contouring function that on correct labelling outputs at least one item. More stringent conditions will be stated later.

Definition 2. Given a nonempty set X, a *concept* c is any[1] subset of X and a *concept class* is a nonempty collection $\mathbf{C} \subseteq 2^X$ of *concepts*.

Definition 3. Given a concept class $\mathbf{C} \subseteq 2^X$, a *sentry* function on $\mathbf{C}$ is a total function

$$\mathbf{S} : \mathbf{C} \cup \{\emptyset, X\} \mapsto 2^X$$

satisfying the following conditions:

(1) the elements of $\mathbf{S}(c)$ are outside c. This means that, for every c in the domain of $\mathbf{S}$,

$$c \cap \mathbf{S}(c) = \emptyset;$$

(2) if $c_1, c_2 \in \mathbf{C}$, $(c_2 \not\subseteq c_1)$ and $c_1 \cup \mathbf{S}(c_1) \subseteq c_2 \cup \mathbf{S}(c_2)$ then $c_2 \cup \mathbf{S}(c_1) \neq \emptyset$;

(3) no $\mathbf{S}' \neq \mathbf{S}$ exists satisfying (2) and having the property that $\mathbf{S}'(c) \subseteq \mathbf{S}(c)$ for each c.

(4) whenever c and c' are such that $c \subset c' \cup \mathbf{S}(c')$ and $c' \cap \mathbf{S}(c) = \emptyset$, then the restriction of $\mathbf{S}$ to $\mathbf{C} - \{c'\}$ is a sentry function of $\mathbf{C} - \{c'\}$.

Notation: $c^+ = c \cup \mathbf{S}(c)$; $c_1 \leqslant c_2$ iff $(c_2 \not\subseteq c_1)$ and $c_1^+ \subseteq c_2^+$; $\mathrm{up}(c) = \{(c' \in \mathbf{C} \mid c \leqslant c')\}$
Terminology: c_2 is *sentinelled* by $\mathbf{S}(c_1)$ iff $c_2 \cap \mathbf{S}(c_1) \neq \emptyset$; $\mathbf{S}(c) =$ the *minimal boundary set* of c upon $\mathbf{S}$.

Remark 2. A given concept class might admit more than one sentry function. The non inclusion of c' in c when $c \leqslant c'$ gives a direction in designing $\mathbf{S}$. The minimality of $\mathbf{S}$ has to be appraised on the whole class $\mathbf{C}$. Condition (4) prevents from building boundary functions which are *unnatural*, in a sense which will be clear in Example 4 below, where some boundary points of some c have the sole role of artificially increasing the elements of c^+ in order to avoid its inclusion in another c'^+. Condition (4) states that this role a can be only considered a side effect of points which are primarily involved in sentinelling some concept.

Example 2. A possible minimal boundary set of an item c of the concept class $\mathbf{H}$ of Example 1 is shown in Fig. 4.

[1] Measurability of c will be taken into account in the next sections.

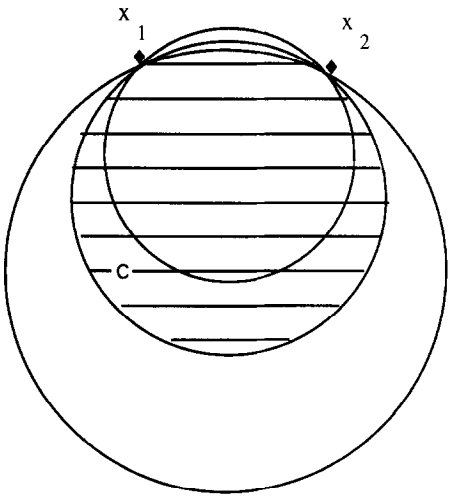

Fig. 4. Two points x_1, x_2 outside c are sufficient to prevent that a larger circle not containing them includes c.

Definition 4. We call *detail* $\boldsymbol{d_C}$ of a concept class $\mathbf{C}$ the supremum of the cardinalities of the minimal boundary sets of its concepts with respect to all possible sentry functions. In symbols,

$$\boldsymbol{d_C} = \sup_{\mathbf{S}, c \in \mathbf{C}} \#\mathbf{S}(c).$$

Example 3. (a) The class of circles of Example 1, has $\boldsymbol{d_C} = 2$.

(b) The class $c_1 = ---$, $c_2 = --++$, $c_3 = +-+$, $c_4 = +++$, where "+" denotes and element x_j belonging to c_i and "−" means an element outside c_i, has $\boldsymbol{d_C} = 2$. A worst case $\mathbf{S}$ is so specified: $\mathbf{S}(c_1) = \{x_1, x_2\}$; $\mathbf{S}(c_2) = \{x_1\}$; $\mathbf{S}(c_3) = \{x_2\}$; $\mathbf{S}(c_4) = \emptyset$. However, a cheaper $\mathbf{S}$ is: $\mathbf{S}(c_1) = \{x_3\}$; $\mathbf{S}(c_2) = \{x_1\}$; $\mathbf{S}(c_3) = \{x_2\}$; $S(c_4) = \emptyset$.

Remark 3. Note that in Example 3(b), in principle, c_2 is not comparable with c_3, but once x_1 is put in $\mathbf{S}(c_2)$ then $c_3 \subset c_2^+$, so that x_2 needs in $\mathbf{S}(c_3)$ to sentinel both c_4 and c_2^+.

From now on we will focus on concept classes like that of our last example. There are two main reasons to do so: They are suitable to put in evidence the combinatorial aspects of the learning problems. The key items of these classes are elements of possibly non vanishing probability.

Let us establish some properties of the sentry functions. These allow us to prove Theorem 1 which is one of the key statements of this paper.

Lemma 1. *Given a concept c in a class $\mathbf{C}$, and a sentry $\mathbf{S}$ over $\mathbf{C}$, for each $c' \in \mathrm{up}(c)$, $\mathbf{S}(c) \cap c' \neq \emptyset$.*

Proof. If $\mathbf{S}(c) \cap c' = \emptyset$ and $\mathbf{S}(c) \cup c \subseteq \mathbf{S}(c') \cup c'$ then $\mathbf{S}(c) \subseteq \mathbf{S}(c') \cap \overline{c'} \cap \overline{c}$, so $c \leqslant c'$ and c' is not sentinelled by $\mathbf{S}(c)$. This contradicts Definition 2. □

Lemma 2. *Adopt the same notation of Lemma* 1. *Further, for each* $x \in \mathbf{S}(c)$, *let* $T_x = \mathbf{S}(c) - \{x\}$ *and let*

$$I_x(c) \equiv \{(c' \in \mathrm{up}(c) \mid c' \text{ is not sentinelled by } T_x)\}.$$

Then we have

(1) $I_x(c)$ *is nonempty.*

(2) *For every* $c' \in I_x(c)$, *we have* $x \in c'$.

Proof. (1) $I_x(c) = \emptyset \Rightarrow \mathbf{S}(c) = T_x$.

(2) If there exists a $c' \in I_x(c)$ such that $x \notin c'$, then c' is not sentinelled by $T_x \cup \{x\}$ and therefore $\mathbf{S}(c)$ is not a minimal boundary set for c. This is a contradiction. □

Corollary 1. *Adopt the notation of Lemma* 1. *Then*

(1) $x \in \mathbf{S}(c)$ *only if there exists some* $c' \in \mathrm{up}(c)$ *such that* $x \in c'$.

(2) *For each* $c \in \mathbf{C}$, *in* $\mathrm{up}(c)$ *there is no* $c' \subseteq c$.

Proof. (1) follows trivially from Lemma 2.

(2) directly follows from Definition 3.2. Lemma 1 shows the necessity of this sentence, because no point x exists which is outside c and inside c'. □

Definition 5. Given a concept class $\mathbf{C}$ and concepts $c, c' \in \mathbf{C}$, we say that c' *does not affect the boundary* of c if for each sentry function $\mathbf{S}$ on $\mathbf{C}$ there exists another sentry function $\mathbf{S}'$ on $\mathbf{C} - \{c'\}$ such that whenever $c'' \neq c'$ is a concept in $\{c\} \cup \mathrm{up}(c)$, we have $\mathbf{S}(c'') = \mathbf{S}'(c'')$.

Corollary 2. *With the same notation of Lemma* 1, *for every* $\mathbf{C}$ *and* $\mathbf{S}$, *for each* $c \in \mathbf{C}$ *such that* X *does not affect its boundary, there exists a* $c' \in \mathbf{C}$ *such that* $c \cup \mathbf{S}(c) \subseteq c'$.

Proof. By induction.

The sentence is trivially true for $\#\mathbf{S}(c) = 1$. Then we assume the sentence true for $\#\mathbf{S}(c) = n$ and we prove that it holds for $\#\mathbf{S}(c) = n + 1$.

Let us denote $\mathbf{S}(c)$ by $X_{n+1} = \{x_1, \ldots, x_{n+1}\}$. Then *any* c'' belonging to a restriction $\mathbf{B}$ of $\mathbf{C}$ on all concepts belonging to $\mathrm{up}(c)$ and containing up to n points $\{x_{i_1}, \ldots, x_{i_n}\}$ of X_{n+1} is sentinelled by a *fixed* subset $\widetilde{X}_n = \{x_{f_1}, \ldots, x_{f_n}\}$ of X_{n+1}.

The sentence trivially holds for any c'' containing at least one point belonging to $\widetilde{X}_n$, from Definition 3. In case c'' contains only $x \in X_{n+1} - \widetilde{X}_n$ from among the points of X_{n+1}, then $c'' \notin \mathrm{up}(c) \cap \mathbf{B}$. In fact, let us assume, ab absurdo, that $c'' \in \mathrm{up}(c)$, then $\widetilde{X}_n \subseteq \mathbf{S}(c'')$. This, by induction hypothesis, happens only if in $\mathbf{B}$ exists a concept c''' containing both x and $\widetilde{X}_n$, and this contradicts the assumption on $\mathbf{B}$.

Thus, all concepts in $\mathrm{up}(c)$ containing at most n points of X_{n+1} are sentinelled by $\widetilde{X}_n$. Therefore, X_{n+1} is a minimal boundary set of c only if a $c' \in \mathrm{up}(c)$ exists such that the whole set X_{n+1} is included in c'. □

Fact 1. *For any concept class* $\mathbf{C}$ *and concept* $c \in \mathbf{C}$,

(1) *no* $c' \notin \mathrm{up}(c)$ *affects the boundary of* c,

(2) *no* $c' \subseteq c$ *affects the boundary of* c,

(3) *a concept* $c' \in \mathrm{up}(c)$ *can affect the boundary of* c *only for the part which is contained in it, i.e. for each* $\mathbf{S}$ *on* $\mathbf{C}$ *and* $c' \in \mathrm{up}(c)$ *a* $\mathbf{S}'$ *exists on* $\mathbf{C}-\{c'\}$ *such that* $\mathbf{S}(c)-\mathbf{S}'(c) \subseteq c'$.

Proof. Let us denote by $A_c \equiv \{(a \in \mathbf{C} \mid c \in \mathrm{up}(a))\}$ the set of the concepts a sharing in their up the concept c, and by $\overline{A_c} \equiv \mathbf{C}-A_c$ its complement.

(i) Let us consider a $c' \notin \mathrm{up}(c)$. From (2) and (4) of Definition 3, for each c, $\mathbf{S}(c)$ takes into account only those c' such that $c \subset c'^{+}$. Now $(c' \notin \mathrm{up}(c)) \Leftrightarrow \neg(c \leqslant c')$. Then, if c' is sentinelled by $\mathbf{S}(c)$, this can happen only if there exists other c'' in $\mathrm{up}(c)$ which have to be sentinelled by the points of $\mathbf{S}(c) \cap c'$, independently of c'. If c' is not sentinelled by $\mathbf{S}(c)$ we could hypothesize that one point of $\mathbf{S}(c)$ is devoted just to make $\neg(c \leqslant c')$. But again this is forbidden by (4) of Definition 3.

Thus, for c, sentinelling points in $\mathbf{C}$ are still sentinelling points in $\mathbf{C}-c'$. Moreover, since $(c' \notin \mathrm{up}(c)\ \&\ (c'' \in \mathrm{up}(c)) \Rightarrow (c' \notin \mathrm{up}(c''))$ we can conclude that $\mathbf{S}'$ of Definition 5 can have $\mathbf{S}'(c'') = \mathbf{S}(c'')$ for each c'' such that $c'' \in (\{a\} \cup \mathrm{up}(a) - \{c'\})$, for some $a \in \overline{A'_c}$.

(ii) Let us assume that some c' and some $\mathbf{S}$ on $\mathbf{C}$ exist such that for each $\mathbf{S}'$ on $\mathbf{C}-\{c'\}$ there exists $c' \in \mathrm{up}(c)$ such that $(\mathbf{S}(c) - \mathbf{S}'(c) - c') \neq \emptyset$.

Now, for each $\mathbf{S}'$ on $\mathbf{C}-\{c'\}$ we can define a $\mathbf{S}''$ on $\mathbf{C}$ just by selecting each $a \in A_{c'}$ and adding to each $\mathbf{S}'(a)$ a point from the set $\mathbf{S}(a) \cap c'$ which certainly makes c' sentinelled by $\mathbf{S}''$, namely the point required to every sentry function, by Lemma 1. Then, let us build $\mathbf{S}'$ by maintaining $\mathbf{S}'(a) = \mathbf{S}(a)$ for each $a \in \overline{A_{c'}}$, from (i), and dropping all unnecessary points from the minimum boundary sets upon $\mathbf{S}$ of the remaining concepts $b \in A_{c'}$. Let us build $\mathbf{S}''$ by adding, if necessary, the point required by S to sentinel c'. Since for each $a, b \in \mathbf{C}$ $(c' \notin \mathrm{up}(a)\ \&\ (c' \in \mathrm{up}(b)) \Rightarrow b \notin \mathrm{up}(a)$, the elimination of points belonging to b might leave $\mathbf{S}(a)$ still unchanged. Then $\mathbf{S}''(a) = \mathbf{S}(a)$ for each $a \in \overline{A_{c'}}$. Let us check what happens on $A_{c'}$. For each $a \in A_{c'}$ with $a \neq c$, $\mathbf{S}''(a) \subseteq \mathbf{S}(a)$, since the added point belongs to $\mathbf{S}(a)$, but $\mathbf{S}''(c) \subset \mathbf{S}(c)$, since, from the assumption, there exists some $y \in \mathbf{S}(c)-\mathbf{S}''(c)-c'$, and this contradicts (3) of Definition 3 on $\mathbf{S}(c)$.

(iii) Putting together the previous points we built a $\mathbf{S}'$ such that c' does not affect the boundary of any $c \in \overline{A_{c'}}$ and affects the boundary of $c \in A_{c'}$ only for the part contained in c'. Thus, (1) and (3) are verified.

(iv) The proof of (2) follows directly from (1) and (2) of Corollary 1. □

Definition 6 (Vapnik [12]). Given a concept class $\mathbf{C}$ and a finite set $Q \subseteq \mathbf{X}$, let $\Pi_{\mathbf{C}}(Q)$ denote the set of all subsets of Q that can be obtained by intersecting Q with a concept in $\mathbf{C}$, i.e. $\Pi_{\mathbf{C}}(Q) = \{(Q \cap c \mid c \in \mathbf{C})\}$. The *Vapnik–Chervonenkis dimension* of $\mathbf{C}$ (shortly, $d_{\mathrm{VC}}(\mathbf{C})$) [13] is the last integer d such that $\max_{(Q \mid \#Q=d)} \#\Pi_{\mathbf{C}}(Q) = 2^d$; if no

such d exists, then $d_{VC}(\mathbf{C})$ is assumed to be infinite. If $\#\Pi_{\mathbf{C}}(Q) = 2^{\#Q}$, then we say that Q is *shattered* by $\mathbf{C}$ [3].

Theorem 1. *For any concept class* $\mathbf{C}$, *concept* $c \in \mathbf{C}$, *and sentry function* $\mathbf{S}$ *on* $\mathbf{C}$, *the set* $\mathbf{S}(c)$ *is shattered by* $\mathbf{C} \cup \{\emptyset, X\}$.

Proof. We will prove this sentence by induction.

We trivially realize that the statement is true for $\#\mathbf{S}(c) = 1$ since $\mathbf{S}(c) \cap c = \emptyset$ and, for each $c' \in \text{up}(c)$, $\mathbf{S}(c) \cap c' \neq \emptyset$. Then we assume the sentence true for $\#\mathbf{S}(c) = n$ and prove that it holds for $\#\mathbf{S}(c) = n + 1$.

Given $\mathbf{C}, \mathbf{S}$ and $c \in \mathbf{C}$ with $\mathbf{S}(c)$ of cardinality $n + 1$, let us consider $X_n = \{x_1, \ldots, x_n\} \subset \mathbf{S}(c)$. Then we extract from $\mathbf{C}$ a new concept class $\mathbf{B} \subseteq \mathbf{C}$ such that there exists $\mathbf{S}'$ on $\mathbf{B}$ such that $X_n = \mathbf{S}'(c)$ and therefore it is shattered by $\mathbf{B}$, hence by $\mathbf{C}$, by assumption. Namely, we prune from $\mathbf{C}$ every concept c' such that either c' does not belong to up(c) or c' contains the point $x_{n+1} \in (\mathbf{S}(c) - X_n)$.

Then we build the new sentry function tightening the past one by removing from the old minimal boundary sets of the remaining concepts all points which do not belong to any $c_i \in \mathbf{B}$ and therefore are useless by Corollary 1.Hence, we certainly remove x_{n+1} from those sets, but, by Fact 1, no removal is due to the pruning of concepts outside up(c). It follows that:

(a) up(c) in $\mathbf{C}$ includes up(c) in $\mathbf{B}$, since $\mathbf{B} \subseteq \text{up}(c)$ in $\mathbf{C}$;

(b) in $\mathbf{B}$ for each c' such that $c \leqslant c'$, c' is sentinelled by X_n. In fact, by Definition 3 in $\mathbf{C}$ we have that for each c' such that $c \leqslant c'$, c' is sentinelled by $\mathbf{S}(c)$. Now the pruning of $\mathbf{C}$ does not include any new concept in up(c) and removes from it concepts sentinelled by x_{n+1}. So each remaining concept has to be sentinelled by X_n.

Therefore, there exists a $\mathbf{S}'$ on $\mathbf{B}$ arising from this procedure such that $\mathbf{S}'(c) \subseteq X_n$. In fact by (b), (1) and (2) of Definition 3 are satisfied by X_n, whilst the conditions of (4) cannot be fulfilled by construction. It remains to check whether a subset of X_n has this property, too.

Let us assume that X_r has the property that $\mathbf{S}'(c) = X_n - X_r$. This means that during the thinning of $\mathbf{S}(c)$ the dropping of x_{n+1} gives rise to the cancellation of a part $q(X_r)$ of X_r from $\mathbf{S}'(c)$. Now, from Corollary 1 we have

$$q(X_r) \subseteq \bigcup_{c_s \in \text{up}(c)} (\mathbf{S}(c) \cap c_s)$$

and from Fact 1.3 it follows that there exists $\mathbf{S}'$ such that the deletion of any $q(X_r)$ must follow the deletion of a concept containing it. But, each $c_s \in \text{up}(c)$ is removed only if $x_{n+1} \in c_s$. Therefore:

$$\text{for each } c_s \in \text{up}(c), \quad q(X_r) \subseteq \mathbf{S}(c) \cap c_s \Rightarrow q(X_r) \cup \{x_{n+1}\} \subseteq \mathbf{S}(c) \cap c_s.$$

So $q(X_r)$ is totally useless as a member of $\mathbf{S}(c)$ in $\mathbf{S}$, and this contradicts the assumptions. It follows that $X_n = \mathbf{S}'(c)$ on $\mathbf{B}$, therefore this set is shattered by the concepts

of $\mathbf{B} \cup \{\emptyset, X\}$. In particular, X does not affect the boundary of c, since $x_{n+1} \in X$. Then X_n is shattered by $\mathbf{B} \cup \{\emptyset\}$, by Corollary 2.

Let us now consider $x_{n+1} \in \mathbf{S}(c) - X_n$. According to the previous procedure, for each $X_n \subset X_{n+1}$ and each subset Q of X_n a concept $b \in \mathbf{B}$ exists such that $Q \subseteq b$ and $(\mathbf{S}(c) - Q) \cap b$ is empty. But there exists a $c' \in \mathbf{C} \cup X$, at least X, such that $X_{n+1} \subseteq c'$. So, adding this partition to all the other partitions containing exactly all the subsets of cardinality up to n of the $n+1$ points of $\mathbf{S}(c)$ we completely shatter $\mathbf{S}(c)$. □

Corollary 3. *For each* $\mathbf{C}$, *detail* $\boldsymbol{d}_{\mathbf{C}}$ *satisfies the inequality* $\boldsymbol{d}_{\mathbf{C}} \leqslant d_{\mathrm{VC}}(\mathbf{C}) + 1$.

Proof. This follows directly from Theorem 1 upon taking into account the possible addition of $\{\emptyset\}$ and X to $\mathbf{C}$. □

Remark 4. In Corollary 6 below we will show that $(d_{\mathrm{VC}}(\mathbf{C}) - 1)/176) \leqslant \boldsymbol{d}_{\mathbf{C}}$. This means that $\boldsymbol{d}_{\mathbf{C}}$ is another, possibly more manageable, way of looking at the same combinatorial property of $\mathbf{C}$.

Remark 5. There do exist concept classes $\mathbf{C}$ such that $\boldsymbol{d}_{\mathbf{C}} < d_{\mathrm{VC}}(\mathbf{C})$. For example, the concept class $\mathbf{C}$,

$$
\begin{aligned}
c_1 &= - - - - - - - - -,\\
c_2 &= + - - - - - - - +,\\
c_3 &= - + - - - - - + -,\\
c_4 &= - - + - - - + - -,\\
c_5 &= - + + - - + - - -,\\
c_6 &= + - + - + - - - -,\\
c_7 &= + + - + - - - - -,\\
c_8 &= + + + + + + + + +.
\end{aligned}
$$

has $d_{\mathrm{VC}}(\mathbf{C}) = 3$ and $\boldsymbol{d}_{\mathbf{C}} = 2$. An example of $\mathbf{S}$ having the largest minimum boundary set is given by: $\mathbf{S}(c_7) = \{x_3\}, \mathbf{S}(c_6) = \{x_2\}, \mathbf{S}(c_5) = \{x_1\}, \mathbf{S}(c_4) = \{x_1\}, \mathbf{S}(c_3) = \{x_1\}, \mathbf{S}(c_2) = \{x_2\}, \mathbf{S}(c_1) = \{x_1, x_2\}$.

Example 4. The following two examples show the relevance of the (4) of Definition 3 in the statement of $\mathbf{S}$. In fact, any $\mathbf{S}$ which does not agree with that condition does not satisfy Theorem 1.

The concept class

$$
\begin{aligned}
c_1 &= - - -,\\
c_2 &= + - -,\\
c_3 &= + + -,\\
c_4 &= + + +.
\end{aligned}
$$

has $d_{\mathbf{C}} = 1$, since the only feasible $\mathbf{S}$ is: $\mathbf{S}(c_3) = \{x_3\}, \mathbf{S}(c_2) = \{x_2\}, \mathbf{S}(c_1) = \{x_1\}$. Actually, $\mathbf{S}(c_1) = \{x_2, x_3\}$ is unfeasible, since the largest concept which is not sentinelled by this minimum boundary set is c_2. On the other hand, if c_2 is cancelled then $\{x_2, x_3\}$ does not remain a minimal boundary set.
Quite different is the concept class:

$$c_1 = - - -,$$
$$c_2 = - + -,$$
$$c_3 = - - +,$$
$$c_4 = + - -,$$
$$c_5 = - + +,$$
$$c_6 = + + -,$$
$$c_7 = + + +.$$

In this case $\mathbf{S}(c_1)$ always consists of a pair of points which constitute an enlargement of c_3 or c_4, but this is unavoidable: as matter of fact, the removal of $\mathbf{S}(c_3)$ or $\mathbf{S}(c_4)$ does not change $\mathbf{S}(c_1)$.

The rest of this section is devoted to a special family of concept classes in terms of contouring functions and of sentry functions.

Definition 7. By a *delta function* $\mathbf{F} : (\boldsymbol{x}, \emptyset) \mapsto \mathbf{F}(\boldsymbol{x}, \emptyset) \subset 2^X$ we mean the restriction of a contouring function $\mathbf{F}$ on the set $\{(\boldsymbol{x}, \emptyset)\}$ of only zero labelled points. This function selects sets of concepts so that all points of $\boldsymbol{x}$ are outside them. We call delta class Δ the union of the concepts belonging to $\mathbf{F}(\boldsymbol{x}, \emptyset)$ with $\boldsymbol{x}$ ranging in X. We denote by Γ the delta class coming from a $\mathbf{G}(\boldsymbol{x}, \emptyset)$ as in Definition 1.

Definition 8. For every precontouring function $\mathbf{G}$, we define $\max(\mathbf{G}(\boldsymbol{x}, \emptyset)) = \{(c \in \mathbf{G}(\boldsymbol{x}, \emptyset) \mid$ there does not exist $c' \in \mathbf{G}(\boldsymbol{x}, \emptyset)$ such that $c \subset c')\}$.

Definition 9. Let ${}^kZ$ be the set of vectors of length k over $Z \subseteq X$. Let B be a set of subsets of X and $\mathbf{B}_Z$ be the quotient set of $\mathbf{B}$ with respect to the equivalence relation on the subsets of X defined by having the same intersection with Z. A delta function $\mathbf{G}(\boldsymbol{x}, \emptyset)$ is *exhaustive* if there exists an n_0 such that, for each subset Y of X and for each $n \geqslant n_0$, the related $\mathbf{G}' : ((\boldsymbol{x}, \emptyset) \mid \boldsymbol{x} \in {}^n(X \cap Y)) \mapsto \mathbf{G}(\boldsymbol{x}, \emptyset)_Y$ is a total function, and $\emptyset$ is the infimum of the set $\{c \in (\max(\mathbf{G}'(\boldsymbol{x}, \emptyset)) \text{ for } \boldsymbol{x} \in {}^n(X \cap Y)\}$. The first n_0 having this property is called the *exhaustiveness size* of $\mathbf{G}$ (in symbols $\mathrm{es}(\mathbf{G})$).

Definition 10. A delta function $\mathbf{G}(\boldsymbol{x}, \emptyset)$ is *congruent* if for any two vectors $\boldsymbol{x}, \boldsymbol{x}'$ over X it never happens that there exist c, c' such that $c \subset c'$, $c \in \mathbf{G}(\boldsymbol{x}', \emptyset)$, $\mathrm{set}(\boldsymbol{x}) \subseteq \overline{c}$, $\mathrm{set}(\boldsymbol{x}) \subset c'$ and $\mathrm{set}(\boldsymbol{x}') \subseteq \overline{c'}$.

Remark 6. The exhaustiveness property of $\mathbf{G}$ allows us to assign a set of subsets of X to every sample $\boldsymbol{x}$ drawn from X, being warranted that at least one sample is the mark

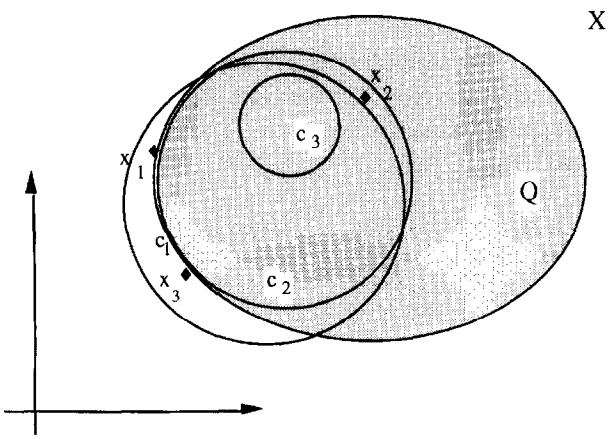

Fig. 5. Exhaustiveness of a delta contouring function. For the delta class $\Gamma = \{X, c_1, c_2, c_3, \emptyset\}$ an exhaustive contouring function $\mathbf{G}$ might attribute c_2 to (x_1, x_2, x_3, x_2) and c_2 to (x_1, x_3, x_3, x_3). Removing Q from X, the corresponding $\mathbf{G}'$ attributes $\emptyset$ to (x_1, x_3, x_3, x_3).

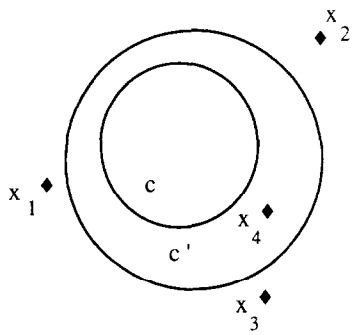

Fig. 6. Congruency of a delta contouring function. It is unnatural and misleading that $c \in \mathbf{G}(x_1, x_2, x_3, x_2), \emptyset)$ when x_4 is available to distinguish c from c'.

of the set $\emptyset$; the quotient set $\mathbf{G}(\boldsymbol{x}, \emptyset)_Y$ is introduced to make this property holds even when some subset $Q = \overline{Y}$ of X has probability measure $P(Q) = 0$ (see Fig. 5). This property is feasible as shown in Fact 2.1 below.

Congruence is a technical expedient to include in $\mathbf{G}(\boldsymbol{x}, \emptyset)$ the set of the largest subsets of X consistent with $\boldsymbol{x}$ (see Fig. 6). Actually, it may happen that the contouring function which we employ is not congruent, but its congruent extension allows us to easily manage some worst-case properties of this function.

Fact 2. (1) *For any concept class* $\mathbf{C}$ *on a finite set* X, *there exists a consistent congruent exhaustive contouring function* $\mathbf{F}(\boldsymbol{x}, \emptyset)$ *with* $\Delta = \mathbf{C}$ *and such that* $\mathbf{F}$ *has exhaustiveness size less or equal the number of elements of* X, *in symbols* $\mathrm{es}(\mathbf{F}) \leqslant \#X$.

(2) *For any* $\mathbf{F}$ *there exists a congruent* $\mathbf{F}'$ *such that, for any* x *and* $c \in \mathbf{F}(\boldsymbol{x}, \emptyset)$, *there exists a* $c' \in \mathbf{F}'(\boldsymbol{x}, \emptyset)$ *such that* $c \subseteq c'$. *If* $\mathbf{F}$ *is exhaustive, then so is* $\mathbf{F}'$.

(3) *If* $\mathbf{F}$ *is congruent then for each* $\boldsymbol{x}$ *we have* $\mathbf{F}(\boldsymbol{x}, \emptyset) = \max(\mathbf{F}(\boldsymbol{x}, \emptyset))$.

Proof. (1) Let us define $\mathbf{F}^{-1}(c)$ as the set Q of subsets of X such that we have $\mathbf{F}(\boldsymbol{x},\emptyset)=c$ iff $\text{set}(\boldsymbol{x})\in Q$. If no such Q exists (e.g. in case $\mathbf{F}(\boldsymbol{x},\emptyset)$ depends on the number of replications of a same components in $\boldsymbol{x}$) then $\mathbf{F}^{-1}(c)=\emptyset$.
Assume $\mathbf{F}^{-1}(c)=X-c$. Thus, $\mathbf{F}^{-1}(\emptyset)=X$. The removal of any subset Y from X maintains the intersection of $\mathbf{F}^{-1}(\emptyset)$ with X-Y permanently inconsistent with any concept different from $\emptyset$.
Afterwards, we can test whether a tightening of $\mathbf{F}^{-1}$ has the same property.

(2) Set $\mathbf{F}'=\text{ADAPT}(\mathbf{F})$, where the function ADAPT is so defined:

function ADAPT($\mathbf{F}(\boldsymbol{x},\emptyset)$)
  for each $c',c''\in\mathbf{C}$ such that $c'\subset c''$
    if $\boldsymbol{x}'$ exists such that $c'\in\mathbf{F}(\boldsymbol{x}',\emptyset)$, $\text{set}(\boldsymbol{x})\subseteq\overline{c'}\cap c''$ and $\text{set}(\boldsymbol{x}')\subseteq\overline{c''}$
    then $c'\in\mathbf{F}(\boldsymbol{x},\emptyset)$.
  for each $c\in\mathbf{F}(\boldsymbol{x},\emptyset)$
    if $\boldsymbol{x}',c'$ exist such that $c'\in\mathbf{F}(\boldsymbol{x},\emptyset)$, $c\subset c'$, $\text{set}(\boldsymbol{x}')\subseteq\overline{c}\cap c'$
      then $c\notin\mathbf{F}(\boldsymbol{x},\emptyset)$
return($\mathbf{F}$)
end function

By inspection, it is easy to see that the output of ADAPT is a congruent function satisfying the constraints of (2) and preserving the exhaustiveness of $\mathbf{F}$.

(3) This follows directly from Definitions 8 and 10. □

Example 4. A consistent, congruent and exhaustive $\mathbf{F}(\boldsymbol{x},\emptyset)$ for the class $\mathbf{C}$ of Example 3 is given by

$$\text{set}(\mathbf{F}^{-1}(c_1))=\{\{x_1,x_2,x_3\},\{x_1,x_2\},\{x_1,x_3\},\{x_2,x_3\},\{x_3\}\},$$

$$\text{set}(\mathbf{F}^{-1}(c_2))=\{\{x_1\}\},$$

$$\text{set}(\mathbf{F}^{-1}(c_3))=\{\{x_2\}\},$$

$$\text{set}(\mathbf{F}^{-1}(c_4))=\{\emptyset\},$$

whose exhaustiveness size is 2.
But for the following class:

$$c_1=---,$$

$$c_2=--+,$$

$$c_3=-+-,$$

$$c_4=+--,$$

$$c_5=+++$$

it is evident that the exhaustiveness size for any congruent $\mathbf{F}(\boldsymbol{x},\emptyset)$ is 3. In fact, $\mathbf{F}$ attributes to x the largest set consistent with it. So we need all the three variables x_1,x_2,x_3 for contouring c_1.

Considering delta functions in their twofold role of concept classes and outcomings of delta functions, allows us to bridge the gap between functions selecting concepts and functions providing minimum boundary sets for these concepts.

Lemma 3. *Given a delta class Δ and its congruent contouring function $\mathbf{F}(\boldsymbol{x},\emptyset)$ we have*:

(1) *there exists a sentry function* $\mathbf{S}$ *and a concept* $c \in \Delta$ *such that* $\mathbf{S}(c) \neq \emptyset$;

(2) *for each* $\boldsymbol{x}$ *and* c *such that* $c \in \mathbf{F}(\boldsymbol{x},\emptyset)$ *there exists a sentry function* $\mathbf{S}$ *such that* $\mathbf{S}(c) \subseteq \mathrm{set}(\boldsymbol{x})$.

Proof. First of all we prove that $\mathbf{S}$ exists. Definition 3 is not satisfied if:

(a) there exist, c, c' such that c' is not sentinelled by $\mathbf{S}(c)$ and $c \leqslant c'$.

Now, from Lemma 1, at least one element of $\mathbf{S}(c)$ must belong to $c' - c$. So we build a set $\beta(c) \equiv \bigcup_{c' \in \mathbf{C}} (c' - c)$ consisting of all the points belonging to $c' - c$ for any c'; hence any c' is sentinelled by $\beta(c)$. Note that $\beta(c)$ satisfies (1) of Definition 3 by construction.

(b) A subset of $\beta(c)$ satisfies (2) in Definition 3. Then we replace $\beta(c)$ by this subset according to the prescriptions in (4) of Definition 3. The following situation could occur: during the pruning $Q = \beta(c)$ is nonminimal, but if we remove a given point y we have at least one c' such that (i) $c \subset c'^{+}$, (ii) c' is not sentinelled by $\beta(c)$ but (iii) $\neg(c \leqslant c')$ and (iv) the removal of c' allows us to cancel y from $\beta(c)$. This means that y only severs to falsify $(c \leqslant c')$. Then we can obtain Q' by substituting in Q y with $y' \in (c' - c)$ and, eventually, continue to prune Q'. Point y' always exists, because of Fact 1.2. But the process game can never terminate if the above substitution yields a new element, say c'', satisfying conditions (i)–(iv). This would mean that c'' is sentinelled by the only y in Q, whence c' does not satisfy condition (iv) (i.e. the removal of c' does not allow us to cancel y from Q). Therefore, conditions (i)–(iv) can be always overtaken during the pruning of $\beta(c)$. Points (a) and (b) define a procedure for building a sentry function $\mathbf{S}$. Let us check the properties of this function.

(1) There exists c such that $\mathbf{S}(c) \neq \emptyset$, since X and $\emptyset$ are added to Δ in Definition 3.

(2) Let $c \in \mathbf{F}(\boldsymbol{x},\emptyset)$ and consider the concept class $\Delta^{-} = \Delta - \{c\}$. By (1) a nontrivial sentry function $\mathbf{S}^{-}$ for this class exists.

Now let us add $\{c\}$ to Δ^{-} and extend $\mathbf{S}^{-}$ to Δ such that conditions (1)–(4) of Definition 3 are satisfied. Let us denote by $\mathbf{S}$ this extension. From Fact 1.2, $\mathbf{S}(c)$ does not depend on any $c' \subseteq c$. On the other hand, assume ab absurdo that no $\mathbf{S}$ can be built such that $\mathbf{S}(c) \subseteq \mathrm{set}(\boldsymbol{x})$. This can only happen if for every $\mathbf{S}$ there is some $c'' \neq c$ such that $c \cup \mathbf{S}(c) \subseteq c'' \cup \mathbf{S}(c'')$ and c'' is not sentinelled by $\mathrm{set}(\boldsymbol{x})$. Then $\mathbf{S}(c) - \mathbf{S}(c'') \subseteq c''$, and $(\mathbf{S}(c) - \mathbf{S}(c'') \not\subseteq \mathrm{set}(x))$. Now, if $c \subset c''$ this contradicts the congruency of $\mathbf{F}$. Let us consider the case $(c \not\subseteq c'')$; in building $\mathbf{S}$ we can avoid $c \cup \mathbf{S}(c) \subseteq c'' \cup \mathbf{S}(c'')$ if we avoid $c \subseteq c'' \cup \mathbf{S}(c'')$. This happens if $c - c''$ is not used for building $\mathbf{S}(c'')$. But, from Corollary 1.2, we *need* $c - c'' \subseteq \mathbf{S}(c'')$ only if there exists c''' such that $c \cup c'' \subseteq c'''$. Now, no point of $c - c''$ belongs to $\mathrm{set}(\boldsymbol{x})$ by definition, no point of $c'' - c$

belongs to set($\boldsymbol{x}$) by our absurdum hypothesis. Moreover, no point of set($\boldsymbol{x}$) belongs to $c''' - c''$, for otherwise this point could sentinel c''' as a point either of $\mathbf{S}(c'')$, so removing the need of having $c - c'' \subseteq \mathbf{S}(c'')$, or directly of $\mathbf{S}(c)$, so removing the need of sentinelling c''. Then either c''' or some $\tilde{c}$ such that $c''' \subseteq \tilde{c}$ would belong to $\mathbf{F}(\boldsymbol{x}, \emptyset)$ and this contradicts the congruency of $\mathbf{F}$. □

3. Measuring Δ classes

From now on, we will consider only measurable subsets of X; this means that any collection of concepts, such as concept classes $\mathbf{C}$ or delta classes Δ, is assumed to be included in the σ-algebra $\mathscr{F}$ of a probability space $(X, \mathscr{F}, P)$.

Let us focus on probability measures of the elements of a delta class Δ. For given Δ, let us consider both a contouring function $\mathbf{F}$ and a sentry function $\mathbf{S}$, and select a concept c of Δ. A subset of X containing c and $\mathbf{S}(c)$ has a probability measure no less than that of $c \cup \mathbf{S}(c)$. For appraising $P(c)$, in this section we relate the event: "$P(c) \leqslant \alpha$" with the event "*a* subset of X of measure α contains at least $\#\mathbf{S}(c)$ many random points".

Notation. By $\underline{X}_m$ we mean the sampling algorithm (briefly sampling) which supplies random sample vectors ξ_m consisting of m independent random variables $\xi_1, \xi_2, \ldots, \xi_m$ drawn from a probability space $(X, \mathscr{F}, P)$. We recall that, starting from the above probability space, the sampling algorithm gives raise to the product probability space $(X^{(m)}, \mathscr{F}^{(m)}, P^{(m)})$, where $X^{(m)}$ is the cartesian product of m copies of X and $P^{(m)}$ is canonically determined by the condition $P^{(m)}(\xi_m) = \prod_{i=1}^{m} \xi_i$.

Definition 11. Given a probability space $(X, \mathscr{F}, P)$, a delta class Δ and a random sampled vector $\boldsymbol{x}$, we define *frontiers* σ's of $\boldsymbol{x}$ in two steps, as follows.

We first let $\theta(\boldsymbol{x}) = \sup_{c \in \mathbf{F}(\boldsymbol{x}, \emptyset)} \{P(c)\}$. Then we define $\Phi(x) = \{(\sigma \in \mathbf{F}(\boldsymbol{x}, \emptyset) \mid P(\sigma) = \theta(\boldsymbol{x}))\}$.

Remark 7. For any $\mathbf{F}, \mathbf{S}$ and P, $\sigma \in \max(\mathbf{F}(\boldsymbol{x}, \emptyset))$.

Fact 3. *Given a delta class Δ and probability measure P, let us denote by* $\mathbf{M}$ *the* maximality class $\mathbf{M} = \{(\sigma \mid \sigma \in \Phi(\boldsymbol{x})),$ *for each* $\boldsymbol{x}$ *from* $X\}$. *Then* $\boldsymbol{d}_{\mathbf{M}} \leqslant \boldsymbol{d}_{\mathbf{C}}$.

Proof. Trivial, since $\mathbf{M} \subseteq \mathbf{C}$. □

The rest of the paper relies on the following lemma.

Lemma 4 (Basic Lemma). *Assume we are given*:
- *a nonempty set* X,
- *an exhaustive delta function* $\mathbf{F}$ *which defines a delta class* Δ *of detail* $d_\Delta = \mu$,

– *a probability measure* P, *and*
– *a sampling* $\underline{X}_m$, *with* $m \geqslant \mathrm{es}(\mathbf{F})$.

Consider the family of sets of random sets defined by $\mathbf{F}(\xi_m, \emptyset)$, *where the first argument of* $\mathbf{F}$ *is a random vector of length exactly* m.

Let us denote by U_μ *the random variable giving the probability measure of a* $c \in \mathbf{F}(\xi_m, \emptyset)$. *Further, let us denote by* u_μ *the analogous probability measure of a* $c \in \mathbf{F}(\boldsymbol{x}_m, \emptyset)$, *where* $\boldsymbol{x}_m$ *is now* fixed. *Let*

$$I_\alpha(\mu, m-\mu+1) = 1 - \sum_{i=0}^{\mu-1} \binom{m}{i} \alpha^i (1-\alpha)^{m-i}.$$

Then we have

(1) *For each* $0 < \alpha < 1$, $P^{(m)}(U_\mu \leqslant \alpha) \geqslant I_\alpha(\mu, m-\mu+1)$. (1)

(2) *If, moreover,* $\mathbf{F}$ *is also assumed to be congruent then for each* $0 < \alpha < 1$ *there exists a distribution law such that* $P^{(m)}(U_\mu \leqslant \alpha) \leqslant I_\alpha(1, m)$. (2)

Proof. First of all we observe that, for any sequence $\mathbf{B} = B_1 \subseteq B_2 \subseteq B_3 \subseteq \cdots$ of subsets of X such that the current one contains all the previous ones, the probability measure u_B of $B \in \mathbf{B}$ is a monotonic nondecreasing function of its size σ_B in any euclidean metrics [15] (see Fig. 7). In particular, for each $\mathbf{F}(\boldsymbol{x}_m, \emptyset), c \in \mathbf{F}(\boldsymbol{x}_m, \emptyset)$ and $\mathbf{S}(c)$, let us consider sequences $\mathbf{B}(c^+)$ containing $c \cup \mathbf{S}(c)$ such that no element included in $c \cup \mathbf{S}(c)$ contains any point of $\mathbf{S}(c)$ and call c^+ the pivot of $\mathbf{B}(c^+)$. Moreover, for any given $c \in \mathbf{F}(\boldsymbol{x}_m, \emptyset)$, consider the sequences $U(c^+)$ of the probability measures of the sets in $\mathbf{B}(c^+)$'s, and let α'' denote the infimum of the probabilities $\geqslant \alpha$ belonging to these sequences. Finally, we remember that for any subset b of X (hence for any item of $\mathbf{B}$, too) of measure α, the distribution law of the random variable N_α which counts the number of points in ξ_m falling inside b is given by

$$P^{(m)}(N_\alpha \geqslant \mu) = 1 - \sum_{i=0}^{\mu-1} \binom{m}{i} \alpha^i (1-\alpha)^{m-i}.$$

Now, for any sampled $\boldsymbol{x}_m$, we consider a concept $c_{\boldsymbol{x}m} \in \Phi(\boldsymbol{x}_m)$. Then we refer to the congruent function $\mathbf{F}'$ defined as in Fact 2.2, and choose a $c_{\mathbf{M}_{x_m}} \in \mathbf{F}'(\boldsymbol{x}_m, \emptyset)$ such that $c_{x_m} \subseteq c_{\mathbf{M}_{x_m}}$.

By Lemma 3, for any congruent $\mathbf{F}'$ and α there exists among the above sequences a certain $\mathbf{B}(c^+_{\mathbf{M}_{x_m}})$ – namely one having a $c_{\mathbf{M}_{x_m}} \cup \mathbf{S}(c_{\mathbf{M}_{x_m}})$ such that $\mathbf{S}(c_{\mathbf{M}_{x_m}}) \subseteq \mathrm{set}(\boldsymbol{x}_m)$ as pivot – and an item $B_{\alpha''}$ such that $n_{\alpha''} \geqslant \#\mathbf{S}(c_{\mathbf{M}_{x_m}})$ implies $\mathbf{S}(c_{\mathbf{M}_{x_m}}) \subseteq B_{\alpha''}$, where $n_{\alpha''}$ is the actual number of points of $\boldsymbol{x}_m$ falling inside $B_{\alpha''}$ (remember that by definition no point of $\mathrm{set}(\boldsymbol{x}_m)$ is inside $c_{\mathbf{M}}(\boldsymbol{x}_m)$). This last event happens iff $c^+_{\mathbf{M}_{x_m}} \subseteq B_{\alpha''}$, provided that $\mathbf{F}$ and then $\mathbf{F}'$ is exhaustive. In fact, the exhaustiveness of $\mathbf{F}'$ implies that for any α there exists $\boldsymbol{x}_m$ and $c_{\mathbf{M}_{x_m}}$ such that $P(c_{\mathbf{M}_{x_m}}) \leqslant \alpha$ and, consequently $P(c^+_{\mathbf{M}_{x_m}}) \leqslant \alpha''$.

Summing up, let us call $U^+_\mu = P(c) + P(\mathbf{S}(c)$ where $c \in \mathbf{F}(\xi_m, \emptyset)$, and $\mathrm{U}^+_\mu = P(c_{\mathbf{M}_{\xi_m}}) + P(\mathbf{S}(c_{\mathbf{M}_{\xi_m}}))$ where $\mathbf{S}(c_{\mathbf{M}_{\xi_m}}) \subseteq \mathrm{set}(\xi_m)$ and denote by small u the analogous

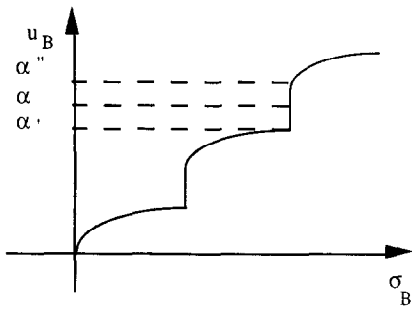

Fig. 7. Rate of growth of the probability measure u_B of B with the enlargement of its size σ_B.

probabilities when ξ_m is substituted by a fixed vector x_m, then:

(a) $(U_\mu \leqslant \alpha) \Leftarrow (U_\mu^+ \leqslant \alpha'')$;

(b) for each $\mathbf{B}(c^+_{\mathbf{M}_{x_m}})$ such that $\mathbf{S}(c_{\mathbf{M}_{x_m}}) \subseteq x_m$, there exists $B_{\alpha''} \in \mathbf{B}(c^+_{\mathbf{M}_{x_m}})$ such that

$$(U_\mu^+ \leqslant \alpha'') \Leftrightarrow (c^+_{\mathbf{M}_{x_m}} \subseteq B_{\alpha''}) \Leftarrow (\mathbf{S}(c_{\mathbf{M}_{x_m}}) \subseteq B_{\alpha''}) \Leftarrow (n_{\alpha''} \geqslant \#\mathbf{S}(c_{\mathbf{M}_{x_m}})).$$

(c) $P^{(m)}(U_\mu \leqslant \alpha) \geqslant P^{(m)}(N_{\alpha''} \geqslant \#\mathbf{S}(c_{\mathbf{M}_{\xi_m}})) \geqslant P^{(m)}(N_\alpha \geqslant \#\mathbf{S}(c_{\mathbf{M}_{\xi_m}})) \geqslant P^{(m)}(N_\alpha \geqslant \mu)$.

To establish inequalities on $P^{(m)}(U_\mu \leqslant \alpha)$ in the other direction we have not so powerful results. But we can rely on a special case which binds from the top the measurability of Δ. To this purpose, let us consider $\mathbf{F}(x_m, \emptyset)$. For any $c^* \in \Delta$, let the point B be in c^* and A outside c^*. In case the distribution law on X is concentrated in A with probability $P(A) = \alpha$ and in B with probability $P(B) = 1 - \alpha$, the crucial point is whether or not $\mathbf{F}(\xi_m, \emptyset)$ contains a concept c containing A or B. Now, $\emptyset$ must belong to $\mathbf{F}(x_m, \emptyset)$ when $\text{set}(x_m) \subseteq \{A, B\}$ because of the exhaustivity of $\mathbf{F}$. Analogously, $\emptyset$ cannot belong to $\mathbf{F}(x_m, \emptyset)$ when $\text{set}(x_m) = \{A\}$ or $\text{set}(x_m) = \{B\}$ by congruency of $\mathbf{F}$. Finally, c will contain B or A only if we don't extract it, by consistency of $\mathbf{F}$. So, under this distribution, for each α, $P^{(m)}(U_\mu < \alpha) = P^{(m)}$ (both A and B belong to $\text{set}(x_m)) = 1 - \alpha^m - (1 - \alpha)^m$. This implies that $P^{(m)}(U_\mu \leqslant \alpha) \leqslant I_\alpha(1, m)$.

Then for each class Δ and for each α there exists a distribution law such that for each $c \in \Delta P^{(m)}(U_\mu \leqslant \alpha) \leqslant I_\alpha(1, m)$. □

Remark 8. Note the different role of the parameters detail d_Δ of the class Δ and exhaustiveness size $\text{es}(\mathbf{F})$ of the contouring function defining the class. The latter one supervises the randomness of the event $P^{(m)}(U_\mu \leqslant \alpha)$, in the sense that it controls the cardinality of the samples allowing us to assume the existence of some c with $U_\mu \leqslant \alpha$. The former parametrizes the probability of this event, by counting the degree of freedom [15] which are burned by the statistics according to the above $\mathbf{B}$ sequences.

Remark 9. Note that in the above proof of the upper bound the hypothesis that $\mathbf{F}$ is consistent represents the best case for having small values of U_μ. So the following corollary holds.

Corollary 4. *Assume we are given a set X, a congruent delta function $\mathbf{G}$ of detail $d_\Gamma = \mu$, and a sampling $\underline{X}_m$. Then for each $0 < \alpha < 1$ there exists a distribution law such that $P^{(m)}(U_\mu \leqslant \alpha) \leqslant I_\alpha(1, m)$.* □

4. PAC learning a concept class

In the following we summarize the basic definitions of PAC learning. Combining the latter with the results of the previous sections we obtain some strengthenings of the basic results on sample complexity.

Notation. We denote by $\underline{X}^c_m$ a sampling which supplies random vectors ξ^c_m consisting of m independent random pair $\xi_1, \chi_c(\xi_1), \xi_2, \chi_c(\xi_2), \ldots, \xi_m, \chi_c(\xi_m)$, with ξ_i drawn from a probability space $(X, \mathscr{F}, P)$, where $\chi_c(\xi)$ is a $\{0, 1\}$-valued random variable such that $(\chi_c(\xi) \mid \xi = x) = \chi_c(x)$ for some concept c on X.

Definition 12. Given a probability space $(X, \mathscr{F}, P)$ and a concept class $\mathbf{C}$, for any given $c \in \mathbf{C}$ we denote by *labelled sample ξ^c_m of size m* the random vector generated by a labelled sampling $\underline{X}^c_m$. By a hypothesis H we mean any statistics on ξ^c_m which defines a random subset of X. For any given x^c_m the value h of H is said to be *consistent with x^c_m* if for every x_i we have $\chi_c(x_i) = \chi_h(x_i)$. We denote by $\mathbf{H} \div c$ the set $\{(h \div c \mid h \in \mathbf{H})\}$, rememeber that $d_{VC}(\mathbf{C} \div c) = d_{VC}(\mathbf{C})$, and define $d_{\mathbf{H},\mathbf{C}} = \sup_{c \in \mathbf{C}} \{d_{\mathbf{H} \div c}\}$.

Definition 13. Given a concept class $\mathbf{C}$ on X, by a *learning algorithm* we mean a function $\mathbf{A} : \{x^c\} \mapsto \{h\}$ such that: for every $0 < \delta, \varepsilon < 1$ there is an integer $m^\circ > 0$ such that for every labelled sampling $\underline{X}^c_m$ with $m \geqslant m^\circ$ and $H = \mathbf{A}(\xi^c_m)$ the probability $P_{\text{error}} = P(\chi_c(\xi) \neq \chi_H(\xi))$ is bounded by the probabilistical inequality: $P^{(m)}(P_{\text{error}} \leqslant \varepsilon) \geqslant 1 - \delta$.

If such a function exists, the class $\mathbf{C}$ is said to be *learnable*, ε and δ are called *accuracy parameters* of the learning algorithm, and the restriction of $\mathbf{A}$ to the set $\{(x^c_m \mid m \geqslant m^\circ)\}$ is said to be a learning algorithm with accuracy parameters ε and δ for $\mathbf{C}$.

Definition 14. Given a concept class $\mathbf{C}$ on X, let us denote by ${}^mZ_c$ the set of labelled samples x^c_m of size m with all the components x_i belonging to Z, and by $\mathbf{B}_z$ as in Definition 9. Then a function $\mathbf{A} : \{x^c_m\} \mapsto \mathbf{C}$ is *strongly surjective* if for each subset Y of X, $\mathbf{A}$ is a surjection from ${}^mY_c$ onto $\mathbf{C}_Y$.

The basic result on the observability of $\mathbf{C}$ is the following.

Theorem 2 (Blumer et al. [4]). *Let $\mathbf{C}$ any concept class with $d_{\text{VC}}(\mathbf{C}) = d$. For any labelled sampling $\underline{X}^c_m$ we have*

(1) *For every probability measure P on X, and any $0 < \delta, \varepsilon < 1$, in case*

$$m \geqslant \max\left\{\frac{4}{\varepsilon}\lg\left(\frac{2}{\delta}\right), \frac{8d}{\varepsilon}\lg\left(\frac{13}{\varepsilon}\right)\right\},$$

every consistent strongly surjective *function* $\mathsf{A} : \{x_m^c\} \mapsto \mathbf{C}$ *is a learning algorithm with accuracy parameters* ε *and* δ *for* $\mathbf{C}$.

(2) *On the other hand, for any* $0 < \varepsilon < \frac{1}{8}$ *and* $0 < \delta < 1/100$, *in case*

$$m < \max\left\{\lg\left(\frac{1}{\varepsilon}\right)\frac{1}{-\lg(1-\varepsilon)}, \frac{d-1}{32\varepsilon}\right\},$$

there exists a probability measure P *on* X *such that no function* $\mathsf{A} : \{\underline{x}_m^c\} \mapsto \mathbf{C}$ *is a learning algorithm with accuracy parameters* ε *and* δ *for* $\mathbf{C}$.

The above result follows from the computation of the probability of selecting, through A, a hypothesis h whose symmetric difference with the unknown $c \in \mathbf{C}$ measures less or equal ε. We can think of the set $\{h \div c\}$ as a delta class Δ coinciding with the union of the sets in the image of the function $\mathbf{F} : (\boldsymbol{x}, \emptyset) \mapsto \mathbf{F}(\boldsymbol{x}, \emptyset) \subset \mathbf{H} \div c \equiv \{h \div c\}$. On the basis of sample $\boldsymbol{x}_m$ of size m, this function selects a set of h's and therefore a set of $h \div c$'s, so that all the points of set$(\boldsymbol{x}_m)$ are outside each one of them. Actually, the differently labelled points of Definition 12 are now all 0-labelled with respect to $h \div c$ when h is consistent (see Remark 1). So we can use the results of the previous sections, which concern any selection h among $\mathbf{F}(\boldsymbol{x}, \emptyset)$ and therefore any learning strategy coded by A. Note that the strong surjectivity of A translates the exhaustivity property of $\mathbf{F}$. In the original formulation of the theorem a companion property is directly charged on the concept class through the notion of *well-behaved concept class* (see the discussion at the end of next section), but in this way occurrences of non exhaustive algorithms on well-behaved classes are not taken into account.

From Lemma 4 we obtain an improved version of Theorem 2.

Theorem 3. *Let* $\mathbf{C}$ *be any concept class with* $\boldsymbol{d}_{\mathbf{C},\mathbf{C}} = \mu$. *For any labelled sampling* $\underline{X}_m^c$ *we have: For every probability measure* P *on* X *and any* $0 < \delta, \varepsilon < 1$, *in case*

$$m \geqslant \max\left\{\frac{2}{\varepsilon}\lg\left(\frac{1}{\delta}\right), \frac{5.5(\mu-1)}{\varepsilon}\right\}$$

every consistent strongly surjective function $\mathsf{A} : \{x_m^c\} \mapsto \mathbf{C}$ *is a learning algorithm with accuracy parameters* ε *and* δ *for* $\mathbf{C}$.

On the other hand, in case

$$m < \lg\left(\frac{1}{\delta}\right)\frac{1}{-\lg(1-\varepsilon)}$$

there exists a distribution law such that no function $\mathsf{A} : \{x_m^c\} \mapsto \mathbf{H}$ *is a learning algorithm with accuracy parameters* ε *and* δ *for* $\mathbf{C}$.

Proof. $\{\mathsf{A}, c\} \subseteq \{\mathbf{G} : (\boldsymbol{x}, \emptyset) \mapsto \mathbf{G}(x, \emptyset) \in 2^{\mathbf{H} \div c}\}$ *in the sense that*, for each A and c there exists a $\mathbf{G}$ such that for each $\boldsymbol{x}_m^c, \mathsf{A}(\boldsymbol{x}_m^c) \div c \subseteq \mathbf{G}(\boldsymbol{x}, \emptyset)$. Moreover, A strongly surjective implies that the related delta function $\mathbf{A} : (\boldsymbol{x}, \emptyset) \mapsto \mathbf{C} \div c$ is exhaustive for every $c \in \mathbf{C}$. Then,

(a) From inequality (1) of the basic lemma,

$$P^{(m)}(U_\mu \leqslant \alpha) \geqslant I_\alpha(\mu, m-\mu+1) = 1 - \sum_{i=0}^{\mu-1} \binom{m}{i} \alpha^i (1-\alpha)^{m-i}$$

implies that $P(U_\mu \leqslant \varepsilon) \geqslant 1-\delta$ is always satisfied if

$$m \geqslant \max\left\{\frac{2}{\varepsilon}\lg\left(\frac{1}{\delta}\right), \frac{5.5(\mu-1)}{\varepsilon}\right\}.$$

(b) From Corollary 4, $P^{(m)}(U_\mu \leqslant \alpha) \leqslant I_\alpha(1,m)$ implies that $P^{(m)}(U_\mu \leqslant \varepsilon) \geqslant 1-\delta$ is satisfied for each distribution law only if

$$m \geqslant \frac{1}{-\lg(1-\varepsilon)}\lg\left(\frac{1}{\delta}\right).$$

Here, with reference to the points A and B of probability $P(A) = \alpha$ and $P(B) = 1-\alpha$ mentioned in the proof of the second part of Lemma 4, the congruence condition on $\mathbf{G}$ is replaced by the applicability of A to each goal concept $c \in \mathbf{C}$ giving rise to symmetric differences belonging to $(\mathbf{G}(\boldsymbol{x}_m, \emptyset) \mid \mathrm{set}(\boldsymbol{x}_m) \subseteq \{A, B\})$, as shown in the proof of Theorem 2 in [4].

Note that in the second part of the proof we do not need $\mathbf{H} = \mathbf{C}$. □

4.1. *What size needs exhaustiveness?*

Consider the following class $\mathbf{C}$:

$c_1 = - - -,$

$c_2 = + - +,$

$c_3 = - + +,$

$c_4 = + + +.$

Note $d_{\mathrm{VC}}(\mathbf{C})$ is 2, but for any choice $\mathsf{A}((x_j, \chi_c(x_j)); j = 1 \ldots m)$, with $m = 1$ there exists a distribution law p such that a c_i exists which will never be inferred by A and the c' selected in its place is such that $p(c' \div c_i) > 0$. For instance, let us assume $\mathsf{A}((x_3, 0)) = c_1$, $\mathsf{A}((x_2, 0)) = c_2$, $\mathsf{A}((x_1, 1)) = c_2$, $\mathsf{A}((x_1, 0)) = c_3$, $\mathsf{A}((x_2, 1)) = c_3$ $\mathsf{A}((x_3, 1)) = c_4$. Now, in case $P(x_3) = 0$, $P(x_1) = \frac{1}{2}$, $P(x_2) = \frac{1}{2}$, c_4 will never be selected, since c_2 or c_3 will always be selected in its place, even if $P(c_1 \div c_2) = P(c_1 \div c_3) = \frac{1}{2}$. By simple inspection we see that no other assignment allows us to avoid this kind of drawback.

Analogously, in [4, Appendix A1] it is shown

Fact 4. *For $X = [0,1]$ and well ordered such that all prefixes of the well ordering are countable, for $\mathbf{C}$ = the set of all suffixes of the well ordering, including the empty set, we have that no learning algorithm A with a finite number of arguments can exactly identify any concept of the class.*

In the absence of evident relations between d_C or $d_{VC}(\mathbf{C})$ and the number of points that A needs to select any c from $\mathbf{C}$ – i.e. to assure the exhaustiveness of the related delta function –, we define, the following measure on $\mathbf{C}$.

Definition 15. Given a learning algorithm A on a concept class $\mathbf{C}$, by a *signature* of $\mathbf{C}$ we mean a function $\mathbf{R} : \mathbf{C} \mapsto \{\mathbf{R}(c)\}$, where $\mathbf{R}(c)$ is a set of labelled vectors $\boldsymbol{r}_c$ of components $(x_i, \chi_c(x_i))$, $i = 1, \ldots, m$ (the length m possibly varying from one vector to another, more than one $\boldsymbol{r}_c$ possibly existing for a single c), such that

(1) (*Exhaustiveness*) With the same notation of Definition 9, for each $(Y \subseteq X)$ and each $c \in \mathbf{C}_Y$, there exists an m and an $\boldsymbol{r}_c$ belonging to ${}^mY$.

(2) (*Meaningfulness*) $\mathsf{A}(\boldsymbol{r}_c) = c$.

(3) (*Minimality*) No subvector of $\boldsymbol{r}_c$ has property (2).

We further define the arity a_{A} of A as follows:

$$a_{\mathsf{A}} = \sup_{r_c, c \in \mathbf{C}} |\boldsymbol{r}_c| \text{ if } |\boldsymbol{r}_c| \text{ is bounded, } \infty \text{ otherwise.}$$

Remark 10. It is not immediate to link a_{A} with parameters characterising $\mathbf{C}$. The value of a_{A} depends both on the representation of $\mathbf{C}$ and on the learning strategy.

As an example, the class of circles and of triangles with edges parallel to three fixed straight lines have both $d_{VC} = 3$ [14] (but the related symmetric differences have $d_\Delta = 2$ and 3, respectively). Now, we can adopt a learning algorithm of arity 2 for the former, when the signature of a circle is given by two points identifying the extremes of its diameter, where an end point is inside or outside the circle depending on its label. For the triangle we must have an arity of at least 3, since not all triangles can be selected by A on two sampled points only.

But, even in case of circles, the learning algorithm might identify circles through their diameter and a tangent passing for a pair of points. Under this learning strategy again we obtain an algorithm of arity >2.

The following example exhibits a learning algorithm whose arity is arbitrarily high for learning a class with $d_{VC} = d_{\mathbf{C}} = 3$.

Example 5. Let X be the set $\mathbf{N}$ of natural numbers and $C = \{\emptyset, \mathbf{N}, c_i, c_{i,j};\ i,j = 1, \ldots\}$, where $c_i = \{i\}$ and $c_{i,j} = \{i,j\}$. Consider the following algorithm:

- $\mathsf{A}((\cdot,0)\ldots\ldots(\cdot,0)) = \emptyset$;
- $\mathsf{A}(\underbrace{(i,1),(i,1),\ldots\ldots,(i,1)}_{h \text{ times}}) = \begin{cases} c_i & \text{if } h \geqslant k, \\ c_{i,i+h} & \text{otherwise;} \end{cases}$
- $\mathsf{A}((i,1)(j,1)) = c_{i,j}$;
- for each $\boldsymbol{x}_m^c$ containing $(i,1)$ h times and some 0-labelled points, let $\mathsf{A}(\boldsymbol{x}_m^c) = \mathsf{A}((i,1),\ldots,(i,1))$, if this concept is consistent with the sample, otherwise $\mathsf{A}(\boldsymbol{x}_m^c) = c_{i,i+h'}$ here h' is the first natural number greater than h such that $c_{i,i+h'}$ is consistent with the sample;
- for each $\boldsymbol{x}_m^c$ containing two points $(i,1),(j,1)$ and some 0-labelled points, $\mathsf{A}(\boldsymbol{x}_m^c) = c_{i,j}$;

– for each x_m^c containing at least three 1-labelled points and some 0-labelled points $\mathsf{A}(x_m^c) = \mathbf{N}$;

Then, the arity of this algorithm is k.

Definition 16. The *arity* $a_{\mathbf{C}}$ and *grain* $\alpha_{\mathbf{C}}$ of a class $\mathbf{C}$ are, respectively, defined to be the sup and the inf of the arity of the learning algorithms for $\mathbf{C}$. In symbols:

$$a_{\mathbf{C}} = \sup_{\mathsf{A}} a_{\mathsf{A}}; \qquad \alpha_{\mathbf{C}} = \inf_{\mathsf{A}} a_{\mathsf{A}}$$

Remark 11. Natarajan [6] introduced a notion of *width* of a concept class which is quite similar to our $a_{\mathbf{C}}$. But that was defined only for Boolean functions with Boolean arguments and was strictly linked to the length of the inputs to these functions. Actually, Natarajan notion of width plays a role which is more similar to our notion of the detail of $\mathbf{C}$. But this role, however, stems from purely counting arguments, which do not apply to continuous X.

Fact 5. *If a signature* $\mathbf{R}$ *is defined for* A, *then its associated delta function* $\mathbf{A} : (x, \emptyset) \mapsto \mathbf{C} \div c$ *is exhaustive for each* $c \in \mathbf{C}$.

Proof. This directly follows from Definition 9 and 15. □

Fact 6. *Given a concept class* $\mathbf{C}$ *and a sentry function* $\mathbf{S}$, *there exists a learning algorithm* A *and a nonexhaustive signature* $\mathbf{R}$ *such that for each concept* c *there exists an* $\mathbf{r}_c$ *such that* $\mathbf{S}(c) \in \mathbf{r}_c$. (*Note that property* (1) *of Definition* 15 *needs not hold*).

Proof (*sketch*). Given $\mathbf{S}(c)$, we add some more point to select c among all the concept sentinelled by $\mathbf{S}(c)$. □

Using Fact 5 we get:

Corollary 5. *For any given concepts class* C, *with* $\mathbf{d}_{\mathbf{C},\mathbf{C}} = \mu$, *for any consistent function* $\mathsf{A} : \{x_m^c\} \mapsto \mathbf{C}$ *with* $a_{\mathsf{A}} = k$, *and for any labelled sampling* $\underline{X}_m^c$, *we have*: *For every probability measure* P *on* X *and any* $0 < \delta, \varepsilon < 1$, *if*

$$m \geqslant \max\left\{\frac{2}{\varepsilon}\lg\left(\frac{1}{\delta}\right), \frac{5.5(\mu - 1)}{\varepsilon}, k\right\}$$

then A *is a learning algorithm with accuracy parameters* ε *and* δ *for* $\mathbf{C}$.

On the other hand, for all $0 < \varepsilon < 1$ *and* $0 < \varepsilon < 1/2k$, *if*

$$m < \max\left\{\lg\left(\frac{1}{\delta}\right)\frac{1}{-\lg(1-\varepsilon)}, k\right\}$$

then there exists a probability measure P *on* X *such that* A *cannot be a learning algorithm with accuracy parameters* ε *and* δ *for* $\mathbf{C}$.

Proof. Sufficiency directly follows from Theorem 3 and Fact 5.

For the converse, consider a consistent A and some c such that $|\boldsymbol{r}_c| = k$, and let us assume that the probability measure is equally distributed on the components of $\boldsymbol{r}_c$, and is 0 otherwise, except possibly for k many points which depend on the adopted learning algorithm. So, the learning problem amounts to guess a proper partition of at most $k + k$ points. Namely, since $\boldsymbol{r}_c$ is a minimum vector, letting $\boldsymbol{q}_r$ be a subvector of $\boldsymbol{r}_c$ obtained by suppressing one component, we have $\mathsf{A}(\boldsymbol{q}_r) = c'$ for some $c' \neq c$. Let $y \in c \div c'$ (possibly $y \notin$ set $(\boldsymbol{r}_c)$) and $p(y) \neq 0$. This means that in learning $\mathsf{A}(\boldsymbol{r}_c)$ through any q_r we meet, with probability 1, a symmetric difference of probability at least $1/(2k)$. □

So, $m \geqslant k$ is necessary to give room to all the hypotheses, while

$$m \geqslant \lg\left(\frac{1}{\delta}\right) \frac{1}{-\lg(1-\varepsilon)}$$

is necessary to maintain low error probabilities.

Remark 12. The lower bound in Corollary 5 is higher than that of Theorem 3, since the condition therein "for any consistent function $\mathsf{A} : \{\boldsymbol{x}_m^c\} \mapsto \mathbf{C}$ there exists a distribution law so that A cannot be a learning algorithm for $\mathbf{C}''$ is sharper than the condition "there exists a distribution law so that no consistent function $\mathsf{A} : \{\boldsymbol{x}_m^c\} \mapsto \mathbf{C}$ is a learning algorithm for $\mathbf{C}''$. Actually, in the absence of any knowledge on the sample distribution law, the first condition appears more realistic, and, however, is the appropriate negation of the learnability conditions.

The second term of the lower bound of Theorem 2 comes from similar considerations, based on the Vapnik–Chervonenkis dimension of C, as the following discussion shows.

Fact 7. *Let* $\mathbf{C}$ *be a class of concepts with* $d_{\mathrm{VC}}(\mathbf{C}) = d$ *and* $\mathrm{vec}(\boldsymbol{x})$ *be a vector whose components are the sole different items of* $\boldsymbol{x}$. *Assume* A *belongs to the family of learning algorithms such that* $\mathsf{A}(\boldsymbol{x}^c) = \mathsf{A}(\mathrm{vec}(\boldsymbol{x}^c))$. *Consider a sampling* $\underline{X}_m^c$. *Then there exists a probability measure* P *on* X *such that for any* $m < d$, $\mathsf{A}(x_m^c)$ *is not exhaustive for some* $c \in \mathbf{C}$.

Proof. Let X_d be a set of points shattered by C. Let the probability distribution P be uniform on these points and 0 elsewhere. From the definition of d_{VC}, we can identify C with 2^{Xd}. So, for any algorithm $\mathsf{A}(\boldsymbol{x}_m^c)$ there exists a c^* which does not belong to the image of the related contouring function. This means that $\mathbf{C} \div c^*$ is not exhaustive. □

If we combine Theorem 2 and Corollary 5, we are unable to tighten definitely the usual gap between the learnability bounds [5]. However, for the class of the consistent algorithms, we have:

Lemma 6. *For any concept class* $\mathbf{C}$ *with* $d_{\mathbf{C},\mathbf{C}} = \mu$ *and* $d_{VC}(\mathbf{C}) = d$, *for any probability measure* P *on* X, *let* ρ *be the ratio between maximum and minimum numbers of examples needed to learn a concept via any consistent function* $\mathsf{A} : \{x_m^c\} \mapsto \mathbf{C}$ *of arity* a_{A}, *with accuracy parameters* $0 < \varepsilon < \min\{\frac{1}{8}, \frac{1}{2}a_{\mathsf{A}}\}$ *and* $0 < \delta < 1/100$. *Then* ρ is bounded by a constant.

Proof. Consider the following inequalities:

$$\max\left\{\frac{2}{\varepsilon}\lg\left(\frac{1}{\delta}\right), \frac{5.5\mu}{\varepsilon}, a_{\mathsf{A}}\right\} > m \geqslant \max\left\{\lg\left(\frac{1}{\delta}\right)\frac{1}{-\lg(1-\varepsilon)}, a_{\mathsf{A}}\right\}$$
$$\geqslant \max\left\{\lg\left(\frac{1}{\delta}\right)\frac{1}{-\lg(1-\varepsilon)}, \frac{d-1}{32\varepsilon}\right\}.$$

We have

- if $a_{\mathsf{A}} = \max\{5.5\mu/\varepsilon, (2/\varepsilon)\lg(1/\delta), a_{\mathsf{A}}\}$, then $\rho \approx 1$.
- If $5.5\mu/\varepsilon = \max\{5.5\mu/\varepsilon, (2/\varepsilon)\lg(1/\delta), a_{\mathsf{A}}\}$, then $\rho < 176$.
- otherwise, ρ, given by the ratio between the first and third term is less than 2.3, since

$$\frac{1}{-\lg(1-\varepsilon)}\lg\left(\frac{1}{\delta}\right) \geqslant \frac{1-\varepsilon}{\varepsilon}\lg\left(\frac{1}{\delta}\right). \qquad \square$$

Corollary 6. $d_{VC}(\mathbf{C}) + 1 \geqslant \boldsymbol{d}_{\mathbf{C},\mathbf{C}} > (d_{VC}(\mathbf{C}) - 1)/176$.

Proof. We obtain from Corollary 3 the inequality between first and second terms, and by direct inspection the inequality between second and third terms. $\square$

If we relax the consistency constraint on A, we might rely on the above parameters a_C and α_C having the following corollary, whose proof is omitted since it follows directly from that of Corollary 5.

Corollary 7. *For any given concept class* C, *with* $\boldsymbol{d}_{\mathbf{C},\mathbf{C}} = \mu$, $a_C = a$, *and* $\alpha_C = \alpha$, *and for any labelled sampling* $\underline{X}_m^c$, *we have*

(1) *For every probability measure* P *on* X *and any* $0 < \delta, \varepsilon < 1$, *if*

$$m \geqslant \max\left\{\frac{2}{\varepsilon}\lg\left(\frac{1}{\delta}\right), \frac{5.5(\mu-1)}{\varepsilon}, a\right\},$$

then every consistent function $\mathsf{A} : \{x_m^c\} \mapsto \mathbf{C}$ *is a learning algorithm with accuracy parameters* ε *and* δ *for* $\mathbf{C}$.

(2) *On the other hand, for all* $0 < \varepsilon < 1$ *and* $0 < \varepsilon < \frac{1}{2}a$, *if*

$$m < \lg\left\{\lg\left(\frac{1}{\delta}\right)\frac{1}{-\lg(1-\varepsilon)}, \alpha\right\}$$

then, for any function $\mathsf{A} : \{x_m^c\} \mapsto \mathbf{C}$ *there exists a probability measure* P *on* X *such that* A *cannot be a learning algorithm with accuracy parameters* ε *and* δ *for* $\mathbf{C}$.

This corollary slightly enlarges the gap between maximum and minimum sample size when some inconsistent hypotheses are allowed.

What is the relevance of a_{A} on the sample complexity of A?
From the bottom, it is easy to to prove the following sentence which casts the parameter for a little part.

Fact 8. *For $d_{VC}(\mathbf{C}) < 17$, the lower bound to the sample complexity does not depend on a_{A}.*

Proof. By direct inspection trying to make the lower bound $(d-1)/32\varepsilon < a_{\mathsf{A}}$ for $\varepsilon < \frac{1}{2}a_{\mathsf{A}}$, where the last bound is required by Corolloary 5. □

On the other direction, it remains an interesting problem for us to identify families of learning tasks and algorithms whose upper bound on sample complexity relies just on the arity of the algorithm.

Fact 4 exhibits a problem which is unlearnable just because the grain of the concept class (see Definition 16) is infinite. This comes, of course, from an inner property of the concept class which characterizes the class as non *well-behaved* [4]. However, in principle, we might think of a nonspecious concept class and a very costly learning function A, whose signatures require, just to mention ε approximations of the goal concept, sample sizes even larger than the usual upper bounds. As matter of fact we did not succeed in finding a similar function. On the contrary, we checked that the algorithm of Example 5 is well ruled by usual upper bounds, in spite of the arbitrariness of its arity.

Where the arity is ineffective we can extend the narrowing of the gap on sample complexity to the wider range of parameters, and remove the constraint of consistency of the learning algorithm.

Corollary 8. *For any concept class $\mathbf{C}$ on X, with $d_{VC}(\mathbf{C}) \geqslant 17$, and any probability measure P on X, the ratio ρ between maximum and minimum numbers of examples needed to learn $\mathbf{C}$ with accuracy parameters $0 < \varepsilon < \frac{1}{8}$, $0 < \delta < 1/100$ is bounded by a constant.*

Proof. It follows from Lemma 6 and Fact 8. □

4.2. The lower cost of subsymbolic learning

In this section we do not state any theorem or lemma. Rather, we just try to adumbrate some realistic management of the new complexity indices and to understand why subsymbolic learning [7] is cheaper.

We start from the prejudice that focusing on the boundary sets is much convenient to subsymbolic learning, whereas using the Vapnik–Chervonenkis dimension is mainly appropriate to the symbolic one.

Suppose we want to learn a formula c which we know, for instance, belongs to the class of k-term-DNF formulas. Having such piece of information, we take this class or some wider one as our hypothesis class. Then we compute its Vapnik–Chervonenkis dimension and state upper and lower bound to the sample size.

By ignoring or disregarding the information about the concept class, we can design a neural network and a learning algorithm for this network. It is commonly claimed that here the hypothesis class coincides with the set of all functions computable by the network on varying the parameters, and that this class must contain that of the target function, in this case the class of k-term-DNF formulas.

Actually, what we want to learn is a given formula under a feasible distribution law, and not an arbitrary one. What we select is a set of network parameters from among those encountered during the training state, and not among all the possible parameters. This reduces the dimensions of the learning problem, since it loses the VC dimension of the (class of all the possible functions computable by the) network. Moreover, the detail of the actual class of hypotheses and the arity of the learning algorithm might be estimated during the training process.

Consider a usual subsymbolic learning procedure which evolves as follows:

Take a neural network and train it on a growing number of labelled examples. At the end of each step test the trained network on N further examples. Stop the training whenever all the labels of the last test sample coincide with the output of the network. Otherwise, continue estimating the size of the next training set and retraining the network.

We point out here two elementary facts:

(i) The event "the subset of the sample space where the label of the samples coincides with the output of the network measures at least $1-\varepsilon$" is equivalent to the event "the probability measure of the symmetric difference between the hypothesis supplied by the network and c is $\leqslant\varepsilon$". Both events imply the event "a signature of a concept c_ε whose symmetrical difference measure with c is $\leqslant\varepsilon$ was completely extracted during the training".

(ii) At the completion of each training step, the hypothesis that the error ε of point (i) is less or equal to a given ε_0 is wrongly accepted by the test sample, against the alternative of $\varepsilon > \varepsilon_0$, with probability at most $(1-\varepsilon_0)^N$.

So we have a possibly never ending learning algorithm whose correctness is tested directly basing on the output performance. Namely, when the algorithm stops we have an error smaller than ε_0 with confidence at least $(1-\varepsilon_0)^N$.

Therefore, the knowledge of the arity and detail parameters is only used to size the next training set. We might determine an approximate value of this size through broad estimates of the arity a_c and the detail μ. Loosely speaking in the light of Fact 6 and point (ii) above, we assume the former as an upper bound on the second. Since the maximum cardinality of the signatures of hypotheses encountered by the learning algorithm is a growing function of the size of the growing training set, we have that this maximum is a lower bound on a_c which can be assumed as a rough estimate of both a_C and μ.

We base this estimate, as well as the expectation of small values of a_c on the following conjecture: During its training story, the network should naturally optimize the shape of these hypotheses (i.e. the values of the net parameters) searching for the minimal arity and detail of a set H_ε of hypotheses which assures an ε-cover of c, with a given probability $1-\delta$. The learning algorithm is the same for each $c \in \mathbf{C}$, but the optimization of the hypothesis complexity strictly depends on: (i) the goal c, i.e. on the net architecture for guessing c, (ii) the past training story or the initialization of the parameters of the net, and (iii) the accuracy parameters which highly condition the evolution of the net parameters during the training. Thus we might view a subsymbolic (neural) learning algorithm as a collection of algorithms, using different classes of hypotheses for each learning target. We might rely on the claimed capability of the neural network for locally lowering the inherent nonuniform complexity of the global class of hypotheses on $\mathbf{C}$, when the network is trained for the more "natural" learnings jobs.

Of course, we can always design a ghastly distribution law which, as in the proof of Corollary 4 and Theorem 3, dumps the above optimizing mechanism. But these artificial distributions are rarely met in real learning problems.

5. Conclusions

General results on learnability usually refer to the families of functions which (i) are consistent with the labelled sample and (ii) include the concept class, as a minimal requisite for the output of the learning algorithm.

In this paper we have examined the consequences of these constraints on the structural properties of the algorithm and we have found out that:

(1) Among the arguments of the learning algorithm A a key role is played by two families of minimal sets of points: those which univocally determine the hypotheses computed by A (signature), and those which prevent A from selecting a hypothesis which properly includes the one determined by the former points (minimum boundary sets).

(2) A learning algorithm must be exhaustive, in the sense that it must be able to output all the concepts of the class to be learnt, modulo equivalence relations induced by the distribution law on the support of the concepts. This obvious requisite might bind the sample size necessary to learn.

(3) Starting from the sets of Point (1), and taking into account the bounds of Point (2), using nonparametric probabilistic techniques, we can show that the ratio between upper and lower bounds on sample complexity is generally bounded by a constant for severe values of the accuracy parameters.

(4) Subsymoblic learning might afford estimates of the complexity of the learning task, and use hypothesis classes of lowest sample complexity.

We dealt with statistics whose degrees of freedom are reduced by the cardinality of the involved minimum boundary sets (see Lemma 4). So the management of the

degrees of freedom, which is so crucial in experiment design, finds here a connection with the complexity of the relations with which we link the sample points. From a dual viewpoint, we give a seed for dealing with noncompletely independent sampled items, in the line of [1, 2].

References

[1] B. Apolloni and G. Mauri, A unified approach to learnability, in: G. Ausiello, D. Bovet and R. Petreschi, eds., in: *Proc. 1st Italian Conf. on Algorithms and Complexity* (World Scientific, Singapore, 1990) 199–217.

[2] B. Apolloni and S. Chiaravalli, PAC Learning is easier than expected, in: L. Di Pace, ed., *Proc. III GAI*IA workshop on Computational Learning* (Roma, 1992) 253–260.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth, Classifying learnable geometrical concepts with the Vapnik–Chervonenkis dimension, in: *Proc 18th ACM Symp. on Theory of Computing* (ACM, Berkeley, CA, 1986) 273–282.

[4] A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, *J. ACM* **36** (1989) 929–965.

[5] A. Ehrenfeucht, D. Haussler, M. Kearns and L.G. Valiant, A general lower-bound on the number of examples needed for learning, *Inform. Comput.* **82** (1988) 247–251.

[6] B.K. Natarajan, On learning boolean functions, in: *Proc. 19th ACM Symp. Theory of Computing* (Ass. Comp. Mach, New York, 1987) 285–295.

[7] D.E. Rumelhart, J.L. McLelland and the PDP research group, *Parallel Distributed Processing* (MIT Press, Cambridge, 1986).

[8] A. Solomonof, A formal theory of inductive inference, *Inform. and Control* **7** (1964) 1–22 and 224–254.

[9] J.W. Tukey, Nonparametric estimation II. Statistical equivalent blocks and tolerance regions – The continuous case, *Ann. Math. Statest.* **18** (1947) 529–539.

[10] J.W. Tukey, Nonparametric estimation III. Statistical equivalent blocks and multivariate tolerance regions – The discontinuous case, *Ann. Math. Statist.* **19** (1948) 30–39.

[11] L.G. Valiant, A theory of the learnable, *Comm. ACM* **27** (1984) 1134–1142.

[12] V.V. Vapnik, *Estimation of Dependence Based on Empirical Data* (Springer, New York, 1982).

[13] V.V. Vapnik and A.Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory Prob. Appl.* **10** (1971) 264–280.

[14] R.S. Wenecour and R.M. Dudley, Some special Vapnick–Chervonenkis classes, *Discrete Math.* **33** (1981) 313–318.

[15] S.S. Wilks, *Mathematical Statistics* (J. Wiley, New York, 1962).